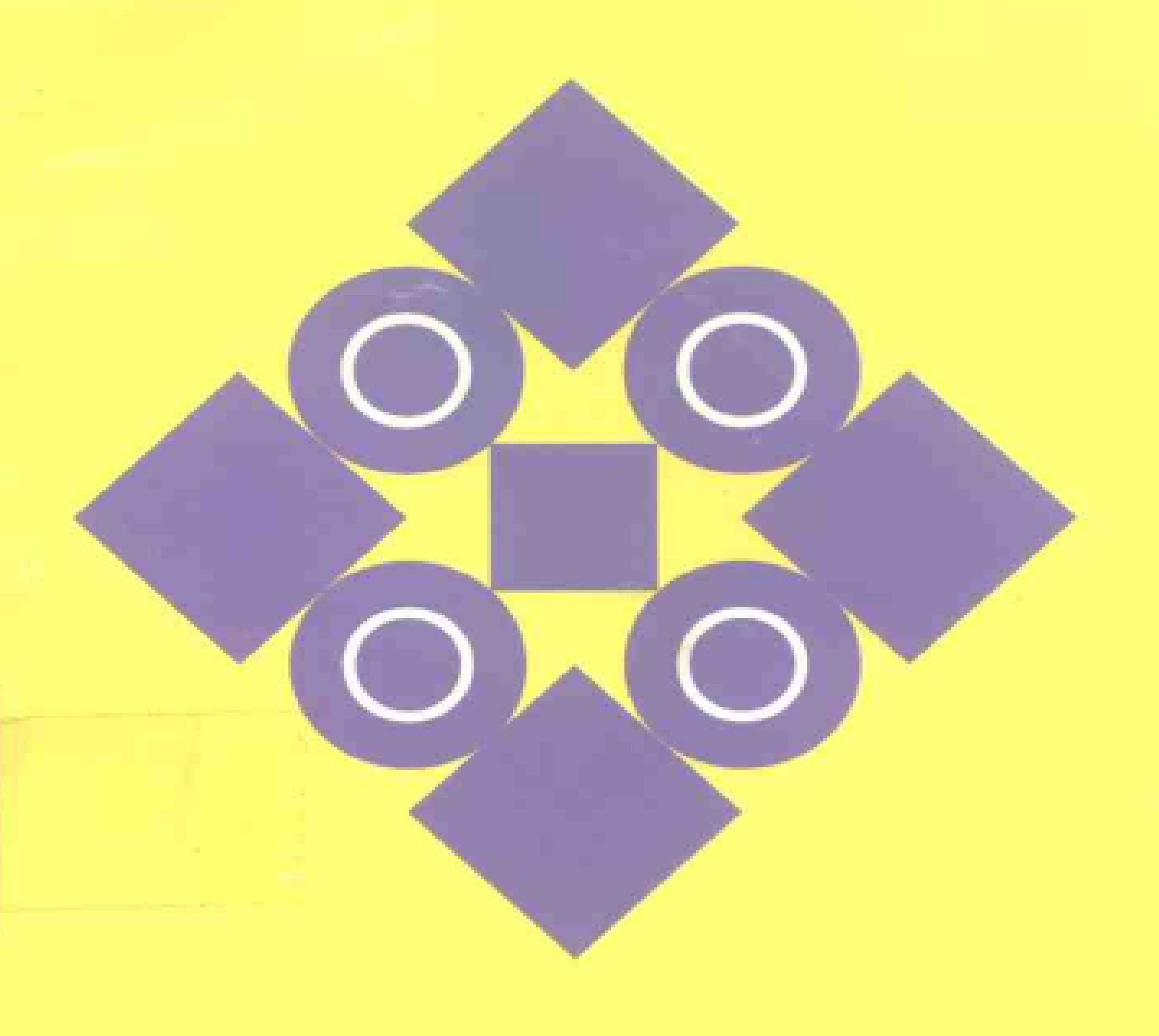
当代学术思潮译丛

计量史学方法导论

著者／[英] 罗德里克·弗拉德
译者／王小宽　校者／袁　宁

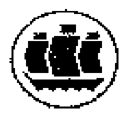


上海译文出版社

当代人文思潮译丛

计量史学方法导论

著者／[英] 罗德里克·弗拉德
译者／王小宽
校者／袁 宁



上海译文出版社

Roderick Floud

AN INTRODUCTION TO QUANTITATIVE METHODS FOR HISTORIANS

Methuen & Co., London and New York, Second edition, 1988
根据梅休因出版公司 1988 年第 2 版译出
Chinese Simplified Character Translation

图字：09-1997-031 号

计量史学方法导论

[英] 罗德里克·弗拉德 著
王小宽 译
袁 宁 校

上海译文出版社出版、发行
上海延安中路 955 弄 14 号
全国新华书店经销
昆山市亭林印刷总厂印刷

开本 850×1168 1/32 印张 8 插页 2 字数 168,000
1997 年 7 月第 1 版 1997 年 7 月第 1 次印刷
印数：00,001—10,000 册
ISBN 7-5327-2036-5/K·064
定价：11.70 元

译者的话

一

计量史学（Quantitative History，也称为定量史学或数量史学）并不是一个严谨的概念。从一般的意义上讲，它是对所有有意识地、有系统地采用数学方法和统计学方法从事历史研究工作的总称，其主要特征为定量分析，以区别传统史学中以描述为主的定性分析。计量史学本世纪上半叶至50年代始于法国和美国，继而扩展到西欧、苏联、日本、拉美等国家。特别是60年代以后，电子计算机的广泛应用，极大地推动了历史学研究中的计量化进程。计量史学的研究领域也从最初的人口史、经济史扩大到社会史、政治史、文化史、军事史等方面；应用计量方法的历史学家日益增多，有关计量史学的专业刊物大量涌现，这方面的论文和专著更是层出不穷（其中亦不乏惊世之作）。至70年代中期，计量史学已成为国际史学研究中最庞大的流派，发展的速度的确相当快。

人类进入了文明社会以后，就再也没有离开过数字，很难想象抛弃了量的概念人们将如何生活。事实上，量的概念早已深入到人类生活的所有方面，构成了人类社会存在的不可缺少的一个组成部分，也

可以说量的概念已溶入人类思维的潜意识之中。自古以来，人类在不断地探索量与量之间的抽象关系而逐步发展起来的数学，以及17世纪以后作为数学的一个分支而发展起来的统计学，对人类认识自然、改造自然起到了巨大的推动作用，数学和统计学也成为众多的自然科学学科赖以存在的基础。既然量在人类生活中的地位如此重要，那么我们在认识自身社会和追溯历史的时候也就不可能将量的概念置之度外。从历史研究的角度讲，没有量就很难准确地描述历史现象，解释历史过程，分析历史的因果关系。从几千年前先人们在撰写历史著作中引用几个数字，到本世纪50年代蓬勃兴起的、系统地应用复杂的数学和统计学方法研究历史的计量史学，中间经历了一个漫长的过程，量在历史著述中的应用无论从深度到广度上都有了质的飞跃。今天，我们可以毫不夸张地讲，定性分析与定量分析的有机结合将是科学地、真实地记录和分析历史的唯一途径。早在一百多年以前马克思就曾表述过这样一种看法，即“一门科学只有在成功地运用数学时，才算达到了真正完善的地步”。①

① 见保尔·拉法格：《摩尔和将军——回忆马克思和恩格斯》，人民出版社1982年版，第95页。

如果从威廉·配第的《政治算术》算起，那么将数学引入社会科学的研究领域甚至可以上推到17世纪。自此以后，随着数学本身(及统计学)的日益完善和社会科学诸学科的日益成熟，这种结合愈来愈紧密。正如列宁在1914年所说："从自然科学奔向社会科学的强大潮流，不仅在配第时代存在，在马克思时代也是存在的。在20世纪这个潮流是同样强大，甚至可以说是更加强大了。"① 1971年哈佛大学的卡尔·多伊奇发表过一项研究报告，其中详细地列举了自1900—1965年全世界的62项社会科学方面的重大进展，并得出如下的结论："定量的问题或发现(或者兼有)占全部重大进展的三分之二，占1930年以来重大进展的六分之五。"② 历史学家当然不可能对这些相邻社会科学领域的进展及发展趋势无动于衷，因为这些成就不仅提高了人们对当代社会发展过程的认识能力，而且很自然地向历史学家提出了这样的问题：显示这些变化过程的历史记录如何呢？当然，这里所要求的历史记录，是能够与上述社会科学所应用的计量方法所得出的结论相对应的精确数

① 《列宁全集》，第20卷，第189页，人民出版社1958年版。

② 转引自丹尼尔·贝尔：《第二次世界大战以来的社会科学》，中国社会科学院情报研究所，1982年版，第2页。

据，而不是传统史学中惯有的描述。显而易见这不是一个传统史学所能够解决的问题，历史学面临着严峻的挑战。历史学家在思考并试图解答这些问题的同时，社会科学其他学科的各种新概念及计量研究方法也就自然而然地渗透到历史研究之中。

二

将定量分析引入历史学无疑增强了历史学家解释和分析历史的综合能力，这主要体现在三个方面。

第一，计量化使历史学研究的对象从传统的、以个人和事件为中心的政治史向以大众和过程为主体的总体史或综合史的转移成为可能，并开辟了史学研究的新领域。本世纪30年代，由M.布洛赫和L.费弗尔开创的法国年鉴学派在西方史学界率先将研究重心转向以经济-社会为核心的整体史，即注重社会和经济活动的宏观过程以及各种较大社会群体的特征和历史演变。这种转变并不必然要求计量分析，但如果历史学家沿着这条路走下去，他就不可避免地要涉及体现那些过程和群体的大量数量型资料，因而也就会自然地把计量作为分析资料、阐述观点、概括结论的主要手段之一。这一点在年鉴学派

第二代身上表现得尤为突出，从F. 布罗代尔的"长时段"到E. 拉杜里的"定量事态史"，统计方法都已成为一种主要的治学方法。50—70年代在美国风靡一时的各种所谓新经济史、新政治史、新社会史等等，其所以冠以"新"字，就在于它们大量借用了与以往的历史研究完全不同的其他社会科学相应学科的概念和方法，而在这些方法中计量方法占了绝大的比重。60年代中叶英国出现的人口统计史及70年代在联邦德国兴起的"社会历史学派"无一不以计量方法为其主要特征。

第二，计量化为历史研究开辟了许多过去不为人重视或不曾很好利用的历史资料新领域。比如，从18世纪后半叶开始西方世界进入了所谓"统计时代"。这一时期正值欧美资本主义制度的确立和发展时期，为了适应资本主义社会化大生产和世界市场范围大规模经济的需要，从18世纪起欧洲各国相继设立了专门的统计机构，负责搜集各种统计资料，定期或不定期地举行人口、工业、农业、贸易、交通等方面的调查，出版或定期公布这些材料。到20世纪中叶，这类统计资料的积累历经200余年，已达到了相当可观的规模。计量方法为系统地、详尽地发掘和利用这些宝贵资料，研究资本主义形成时期至今的这

段历史提供了一个非常有力的工具。当然，举这个例子并不意味着计量史学研究仅局限于“统计时代”以后的近现代史范围，如许多研究欧洲中世纪的学者通过计量手段使记载着教民出生、婚姻、死亡等内容的教区登记簿展现出新的重要价值。历史研究中的计量化还推动了收集和整理数量型历史资料的工作。早在1962年美国的一百多所研究机构和大学就建立了一个计算机可读的美国政治史资料库，它储存了有史以来有关美国总统、州长和议会选举的全部档案材料以及相关的各个时期的经济、社会、人口资料，以后这个资料库又将收集的范围扩大到130多个国家和地区。从此以后各国学者又陆续建立了许多有关一个村庄、城市、地区及国家的规模不等的数据库。这些工作为计量史学的进一步发展打下了良好的基础。

第三，计量化使历史学趋于严谨、精确。对于传统史学来说，无论是定性分析还是描述事实都要以文字的形式来表达，而用文字语言解释历史却有一个难以克服的技术上的问题，这就是文字语言的多义性和模糊性。细心观察一下就不难发现日常生活中所使用的文字语言的许多词（特别是涉及到理念及价值判断的词语），在不同的时间、地点、场

合，在发出和接收它们的人们之间都会存在着某种差异。由于历史研究是当代人与过去的对话，这个问题更为突出，历史研究中的许多争论问题，与其说是由于观点、立场上的分歧，不如说是由于对某些关键性的概念及词语的不同理解所造成的。尽管人们早已认识到了这个问题，但由于历史学研究的对象是具有主观意识的人的社会活动而不是无意识的自然客体，历史学至今未能像自然科学或社会科学中的经济学、人口学和社会学那样形成一套严谨而精确的专业性语言。然而，我们还应该认识到既然人类社会生活中存在着大量的数量关系，那么用数学语言来表述某些社会历史现象或者将一些非数量型的社会历史现象用某种数量形式表述出来，无论是逻辑上还是方法上都是可行的（这并不是说人类社会的所有方面都可以转换成数量形式）。事实上，在上述社会科学的诸领域中人们早已将社会生活的某些同质或同类的要素非人格化，即将人类的活动转化成抽象的数字，再运用数学语言对其进行分析研究。同文字语言相比，数学语言有这样一些特点：其一是准确性。在涉及到现象的规模、速度、程度、范围等方面，数学语言的优势是非常明显的。我们说定量分析比定性分析更准确，实际上就是因为用来

表达定量分析之结果的数学语言较之定性分析的文字语言具有更高的精确度的缘故；其二是通用性。数学语言不受时间、地点、语种和不同体制等人为因素的限制。它的内涵十分清晰明确，不存在意义的外延问题；其三是可比性。这一特点在历史比较研究中很突出，无论是纵向比较还是横向比较，数学语言比文字语言更容易建立起一个绝对的比较标准；其四是简洁性。数学语言往往仅用若干指标、指数、等式或一组模型就可以说明一个非常复杂的社会现象或社会发展过程。上述这些特点使得采用数学语言的计量方法在评述具有数量特征的社会历史现象时比文字语言更准确、更有力。

三

计量史学近几年才被介绍到国内来，并正在引起越来越多的史学工作者的重视。应该说，计量史学在中国的前景是广阔的。首先，40年来中国史学研究从整体上讲是以马克思主义的历史唯物论为指导的，而从历史认识论的角度看，唯物史观的基本原则不仅与计量方法之间没有相抵之处，相反却为历史研究中应用计量方法提供了广大的天地。关于这

一点法国著名的马克思主义历史学家A.索布尔曾有过一段精辟的论述。[1] 他认为马克思主义历史学与历史社会学有众多的相似之处,所有的政治史,从某种意义上说都是社会史,而对社会史的最终分析,只能是计量史。因为应用于社会史研究的资料必须以精确的有关社会结构、社会变化过程和社会经济事态的知识为前提条件,绝不能是"脱离实际的"或"非社会性的"。这类资料本质上就具有数量特征。再者,从资料方面看,研究中国历史的学者更具有得天独厚的条件,这就是两千多年来中国的历史记录一直未曾中断,而且丰富多样。就拿最适合于计量化的经济史、人口史方面的资料来说,正统的二十五史中大多都有专业记载税赋、田亩、人口及其他经济情况的《食货志》,成为一种定例。这些《食货志》中包含了大量的数量型资料,对于计量史学研究非常珍贵。除此之外,在浩如烟海的历史典籍、文献,特别是数量众多的地方志中数量型资料更是不胜枚举。我们完全有理由期望计量方法将在中国史学研究中发挥更大的作用。

---

① 参见G.伊格尔斯:《欧洲史学新方向》,华夏出版社,1989年版,第168页。

四

《计量史学方法导论》一书初版于1973年，中译本根据1979年的第二版译出。尽管本书作者在引言中声称，这本书不是一部有关统计学的教科书，然而1973年它一问世立即受到国际历史学界的高度评价，在欧美地区特别是英语世界里许多大学的历史系都选用此书为训练学生掌握计量方法的标准教材，更多的历史学家则把它作为涉足计量史学领域的入门书。本书的篇幅虽然不大，却包括了一般计量史学中所使用的一些最基本的统计学方法和技巧。

任何历史学家在从事历史研究时都要首先对历史资料进行某种形式的分类，以适合自己的研究需要，计量分析亦不例外。弗拉德认为对资料的分类是系统研究历史证据的基本要求之一。因此，在第一章里他就提出了对历史资料加以分类的三种类型：定名资料、定序资料和区间资料。在计量分析中不同类型的资料将使用不同的统计方法，对历史资料的分类是否准确，将直接影响到计量分析的准确性。在第二章里，作者叙述了完成对资料的分类以

后如何对此加以进一步的整理以达到统计方法的要求，解释了计量分析中的一些基本概念，如资料集、个案、变量、资料矩阵，等等。第三章介绍了几种简单而又实用的统计方法的计算，以及简化数学运算的一些技巧。第四章和第五章的内容即所谓"描述性统计"方法。由于描述性统计不需要高深的统计数学方面的知识，便于掌握而且应用这类统计方法得出的结论易于理解，因而为大多数计量史学家所采用。其中有频数分布法、图表法，以及利用若干指标来反映总体的基本特征和规律的概括性方法，即各种平均数的计算和应用。第六章为时间数列分析（也称为动态分析）。时间数列分析从数量方面研究历史现象发展变化的趋势和速度，揭示历史现象不同发展阶段的特点和规律，在计量史学中占有重要地位。第七章为相关分析法。它涉及到如何确定两种或两种以上的历史现象之间是否存在着某种关系、其关联的程度如何，以及它们的形式等问题。第八章着重讨论资料的缺失问题。由于种种原因，在大多数情况下历史学家并不能完整地得到他所需要的全部资料，数据的缺失问题在运用计量方法研究历史时显得更为突出。为此，作者专门用一章的篇幅来讲述解决这类问题的方法。在第九章里作者向读

者介绍了电子计算机的基本情况，以及历史学家怎样应用电子计算机进行计量分析。最后，本书还提供了一份较为详尽的参考书目，对于那些想进一步了解计量史学的读者来说，它颇有参考价值。

本书作者罗德里克·弗拉德教授1942年生于英国伦敦，曾先后在伦敦大学和剑桥大学讲授近代史和经济史，除本书外，还曾出版过《经济史论文集》、《1700年以来的英国经济史》等著作。

目次

引言

当我们描述和分析存在于过去或当代的人类社会时，我们不可避免地要使用数字和数量。假如我们要对某一个人作一番充分的描述，那么他的年龄、出生日期、财产、妻子的数目、孩子的数目等等，都是我们必须了解的数量特征。在作这样的了解时，我们把他与其他人进行衡量和比较，是较富还是较穷，较年长还是较年轻，并试图通过这些方法，以及对他的思想和工作的讨论，确定他在其生活的那个社会里的位置。通常我们把行为或思想相类似的人分成各种集团。我们用“中产阶级”、“法国人”、“保守派”这些名词术语来描述这些集团。我们必须以这种方式来进行分类和分组，因为只有这样才能将复杂多样的人类思想和行为变为可以处理的形式。

像年龄、财产、子女的数目这类衡量显然是计量性的。我们只有通过计算他出生以来的年份的数目才能衡量一个人的年龄，只有通过计算他所拥有的一定数目的或以一定数目的货币单位表示的实物或实物价值我们才能衡量他的财产。如果我们在描述生活在过去的人们时使用这类衡量，我们所用的即是计量方法。与此对照，我们在历史研究中所使用的

其他衡量和描述的形式是非计量性的。所描述的是个人或集团的思想或态度;"法西斯主义者"、"文艺复兴时期的人"就是这样的描述。但是当我们从事这些非计量性或定性的描述时常常会发现，只有通过衡量拥有这些观点的或者可以用这类名词加以描述的人们的数目，我们才可以赋予它们一个充分的含义,并估计它们的历史意味。例如"中产阶级"是对社会中一个集团的描述，但是在许多场合它也是对社会中具有特定收入和态度的一定数目的人的描述。如果我们说"中产阶级支持政府",我们意指大多数(如果不是全部)被我们描述为"中产阶级"的人支持政府，而且只有通过计算这些人的数目我们才能够最终证实这一陈述的正确性。很多历史学家所用的定性判断和描述因而蕴含着一种计量的意味，有时这种意味需要明白表示出来。此外，许多对个人或集团行为的描述都含有计量的意味;像"通常"、"一般"、"经常"、"许多"这些词都指数量概念,而且虽然一般我们不会去精确地加以验证,在原则上它们的意味或正确性只能通过计量性衡量来确定。

像其他社会科学家一样，历史学家因而经常和不可避免地应用计量的概念。这个事实并不意味着他们所作的所有陈述都是计量性的，也不意味着他们认为人类行为的所有方面都可以衡量并给予数字。人类及人类中的不同集团和事物的许多侧面都不能以数量来衡量和表达；虽然我们可以测度一个中世纪农民的收入，我们却无法测度他花费这一收入时所取得的乐趣。同样,虽然我们可以测度 15 世纪生产的绒面呢的价格变化，我们决不会知道它们的手感如何。事实上，与其他社会科学家相对比，历史学家在他的测度方法上是尤其有限的;他不能询问他的研究对象有关其幸福或态度的问题,

因此甚至没有希望能像心理学家和社会学家所做的那样，对幸福或政治态度作出相对的衡量。

尽管如此，昔日人类经历的某些方面不能测度这一事实，并非不去测度那些可以为我们理解的那些经验方面的理由。至少，可以测度的方面有助于我们去解释难以测度的方面。像小阿瑟·施莱辛格那种观点，“几乎所有的重大问题之所以为重大，恰恰是因为它们不能以计量作答”，① 忽略了这样的事实，即没有计量答案我们可能就无法解释“重大”问题的证据所在。如果我们确定了某人的收入在增加而不是对他的收入一无所知，那么就更容易解释此人的幸福也在增加。因而即使我们基本上对“定性”比“定量”问题更感兴趣，两者仍是不可分解地联系在一起的。定量问题补充定性问题，定量证据补充定性证据；两者无法相互取代，两者各自也不能以了解整个历史学的研究自命。不论他的兴趣何在，历史学家的一个主要问题是他永远面对着不完备的证据；我们从来就没有足够的证据可以有把握地说我们的解释或描述是完全正确的。如果在贬低计量化的重要性时，历史学家将计量证据排除在外或者将它降到一个附属的地位，那么他就是在进一步缩减他所得到的已经不完备的证据。计量证据几乎肯定不会提供一个全面的答案，但是它很可以提供一部分的答案，而把这部分的答案视若无睹地丢掉，既是浪费也是不负责任。

对计量历史学的另一种更为严重而同样错误的批评是，运用计量方法必然陷入过分简单化，丢失有关过去的信息，将

---

① 小阿瑟·施莱辛格：“人文主义者眼中的经验主义的社会研究”(Arthur Schlesinger jun., ‘The humanist looks at empirical social research’)，载《美国社会学评论》第26卷(1961年12月)，第770页。

个人强行纳入各种类别，以及随之而引起历史的非人性化。当然，分类或综合方法的任何应用，都会将多样化的人类历史经验简单化，因为这就是使用这些方法的目的。没有一个历史学家能够完全穷尽那样的繁复性，而历史学家的心情同任何面对着繁复现象的人的心情一样，不可避免地要去寻求类型和类似性，同时抛弃或忽略许多不适合这些类型的东西。与定性和印象主义的历史学相比较，计量历史学的优点在于它的分类的体系和方法，它所用的假设和所立的类型都是被宣明而清楚的。由于资料的分类和排除是一目了然的，人们没有必要去窥测历史学家的内心或追随他的思路以理解计量历史学。在明确地寻求类型和类似性时，计量历史学家也总不得不承认他是在进行简单化，并叙述他是怎样做的；他不会无意识地删除不利于他的证据项目。计量历史学家因而决不会看不到历史证据的固有的繁复性，他所设计的测度方法正是将这一繁复性约简为能够理解的形式，但也对所用各种类型和平均数中的证据的偏差提供指南。

也有一些质量欠佳的计量历史学，其中证据被强行纳入预定的类别，或者所用的假设与历史事实相反。无论如何，很难说没有不良的计量历史。但可以肯定的是，历史学家所做的大部分（如果不是全部）陈述都是计量性陈述，许多历史证据也是计量性的，并应该用计量方法去分析，而采用计量方法的历史学家应善于利用它们。正如对中世纪手稿的辨识，对启蒙运动时期政治思想的解释需要人们具有经验，受过训练和掌握技巧，处理计量材料也需要人们懂得分析的特定方法和技术。历史学家不能简单地研究一个数字图表而期望立即发现其中的意味；他必须学会抽绎出其中含义的技术，并将此

含义与他所收集到的其他证据联系起来。因此，本书的目的就是帮助计量历史学家善于利用他的材料，并帮助那些阅读他的著作的人们判断他是否这样做了。

下面各章联系它们在历史问题和历史证据中的应用，讨论了一些计量技术。第一章讨论对历史证据进行分类和整理的方法，以便运用本书后面所叙述的方法对它进行分析。第二章里讨论以概括形式来描述证据的方法，而第三章则讲述一些简单的数学技术，它们在证据的分析中很有用处（不需要具有超过简单的中学算术和初等代数的数学知识）。所以，这三章描述了历史学家开始他的分析之前所必须从事的预备性步骤。

第四章和第五章叙述了分析的初步的几个阶段。第四章里讨论用图形和图表形式表述证据的方法，而在第五章里，讨论对集中趋势（平均数）的测度方法和离中测度方法，以补充上述呈现证据的方法。在第六章中，所有这些技术都应用于按年代顺序排列的证据（时间数列），并与一些在时间数列分析中十分重要的技术一起讨论。

在第七章里，应用本书前面所谈到的方法和概念以讨论关于确定两组证据之间关系的存在的统计方法。还讨论相关的概念，并简单介绍一种最有用的统计方法——简单线性回归。

第八章涉及历史资料的一个特殊问题，在传统的统计学教科书中一般没有考虑到它——证据缺失的问题。根据以前几章所讨论的方法，提出了解决证据缺失问题的若干方法，并介绍了抽样的概念。

最后，在第九章里叙述计量分析的工具：电子计算机和

电子计算器。本书前面所讨论的种种方法并不要求使用这些设备，但可以看出，有了它们的帮助，计量分析会变得更简便，更节省时间。

本书既不是一部统计学的教科书，也不是对历史学方法的讨论所作出的一种贡献。它也不能对计量材料的分析中可能出现的所有问题提供答案。近年来，对历史问题进行的计量研究迅速增长。这些研究使用了种种从其他社会科学汲取来的方法，它们的数量如此之多，或在某些事例中如此之复杂，以至无法在本书中一一加以描述。不管怎样，这些研究和方法都具有本书以下各章所解释的一些基本统计技术的共同核心。本书是对历史学家所必需的一些技能的介绍，不论他想要阅读使用统计分析的书籍或论文，还是自己使用计量证据研究政治史、社会史、人口史、经济史，甚或思想史。本书也是计量历史学家最终都可能需要阅读或查询的统计学、计算、计量经济学或数学等其他许多书籍的入门书。

1 历史资料的分类

对历史证据进行系统研究的基本要求之一就是必须把材料分类。根据他的先入之见和所研究的对象，历史学家自然以多种方式对他的材料加以分类。例如，历史学家一般将他们的材料分为第一手证据和第二手证据。第一手证据是在所研究的历史时期所产生的证据，而第二手证据则通常是经过了其他历史学家某种形式的再加工后得到的证据。历史学家所采用的其他分类方案区分文字的和考古的证据，书写的和印刷的证据，或定量的和定性的证据。历史学家还可以采用更为详细的分类方案；例如按照其来源的不同，人们可以将第一手证据分为日记、法律记录、法庭记录、报纸、选举结果、商业记录。

凭借使用种种记录的经验，历史学家逐渐形成了一些准则，使他们能够据以判断出不同类型材料的价值，并帮助他们有效地使用这些不同类型的材料。因此，他们将自己的材料分类，部分是为了可以更容易应用这些准则。因而，例如，J．J．巴格利将1660—1770年的教区记录分为两组：人头税统计表和洗礼、结婚、丧葬的教区记录簿。对这两组加以区分之

后，巴格利认为，人头税统计表对于人口变化并不是一个可靠的指南，而教区记录簿有时会成为可靠的根据。①

使用计量材料的历史学家必须学会不仅根据它们的来源和可靠性，而且还根据它们在多大程度上显示出适用于不同的分析方法来将自己的材料分类。他必须采取的第一步是检查他的资料，即我们将称之为他正在分析的材料，并以有助于他分析的方式对它们进行分类。可以区分出我们能够采用的三种分类类型：定名，定序和区间。我们能否以这三种类型之一将资料分类，完全取决于我们所拥有的信息和证据的数量。

1.1 定名分类

第一种和最简单的分类形式，是我们在日常言语中所用的，即我们把事物通过赋予其名称而区分为属类；它常是计算每一类包括多少事物的第一步。例如，《末日裁判书》②的编纂者在记录1086年肯特郡的威(Wye)采邑时所提供的信息就是定名资料：

有52犁③土地。庄园中有9犁土地，114名佃农有22块边沿地，合17犁。有一座教堂和7名农奴，4座磨坊

---

① J. J. 巴格利：《历史的解释，2：英国史资料，1540年至今》(J. J. Bagley, Historical Interpretation, 2: Sources of English History, 1540 to the Present Day)，哈孟兹沃斯：企鹅图书出版社，1971年，第84—154页。

② 《末日裁判书》(Domesday Book)1086年英国威廉一世颁布的全国土地、财产、牲畜和农民的调查清册。——译者

③ 犁(plough)为英国历史上可耕地面积单位，约合八头牛一年中可耕的土地。——译者

价值23先令8便士，有113英亩的草地和林地，根据林地牧猪权税应缴纳300头猪。①

在这个例子中，《末日裁判书》的编纂者检查了威采邑中的各种物质对象、人、家畜和农业工具，赋予它们名称并得出每一类的总数。

指出对威采邑描述中的若干特征是重要的——这些特征一般都应用于定名分类。第一个特征是，从原则上讲赋予威采邑中每一特性的名称是任意的。如果我们赋予这些特性以拉丁文的名称，一如在原稿中所赋予的那样，或以我们自己的某种新语言的名称，对这一村庄的描述不会有什么不同。所赋予的名称并不重要；只要《末日裁判书》的编纂者和它的阅读者双方都同意以这些名称赋予某些特征，这些名称就是符合要求的。

定名分类的第二个特征是，它并不含有特征排列的次序具有任何特殊的目的之意，并且也没有威采邑中的一个特征比另一个更为重要之意。虽然这些特征是按照所引原文的次序给出的，但如果它们按不同的次序排列，也不会对这一村庄描述的准确性产生什么不同。事实上，《末日裁判书》的编纂者在每一项记录中大致上保持相同的排列次序，这样有利于对采邑间进行比较，但是在每一采邑中，次序是无关紧要的。

正如描述威采邑时那样，定名分类的第三个重要特征是，

① J. J. 巴格利：《历史的解释，1：英国中世纪史资料，1066—1540年》(J. J. Bagley, Historical Interpretation, 1: Sources of English Medieval History, 1066—1540)，哈芒兹沃斯：企鹅图书出版社，1965年，第27页。

项目所列入的各个类是没有联系的或相互排斥的，并且除了是同一采邑中的不同特性之外，它们之间没有任何关系。例如，不可能将牧猪和磨坊相加，并得出威采邑有304牧猪/磨坊的结论，因为牧猪和磨坊这两类彼此分立，不能合计。甚至对于犁的情况，看来似乎有可能推断出这采邑总共有26犁土地——9犁在庄园内，17犁在庄园外——我们也不能打破这一准则；事实上我们不是在将庄园土地与非庄园土地这两个不同的类相加，而是建立一个包括这两者的新的类——所有类型的土地。像这种将各类编组运算通常是可能的，但应该认识到在编组中我们并没有将两个类相加，而是建立了一个新的类。

1.2 定序分类

在很多事例中，我们所有的信息数量或者我们关于资料愿作出的假设数量，使我们可以比仅仅列举我们感兴趣的某些事物的特性稍进一步。可以在我们所用的类上施加某种次序，而说组成一个类的项目比组成另一个类的项目要大些，老些，小些或丰富些。假如对已经建立起来的各类之间的关系能够作出这样的陈述，这种资料就被认为是定序类型的资料。

在历史著作中经常碰到的定序分类的一个例子就是社会地位。例如，1688年格雷戈里·金为英国人口中的社会各阶层编制出一张一览表，以及对每一社会阶层的家庭数目的估计。表1.1给出的就是从这个一览表中摘录的显示在他的26个社会阶层中的前13个，即社会结构的上层一半部分的情

况。在对社会阶层的这种分类中，类的编纂者不仅列举了各种类以及象定名分类那样给出归属各类的项目数，而且他还感到有可能对一类与另一类之间的关系作出陈述。格雷戈里·金不仅计算出世俗贵族和精神贵族的家庭数目，并且判断出前者多于后者。

在定名分类中次序的排列无关紧要，即使将它们打乱也不会有什么不同，而对于定序分类，次序正如"按次序"这个形容词所意指的那样，是非常重要的。假如我们将格雷戈里·金表中的各类打乱，并按一个不同的次序排列它们，那么我们将失去他这个一览表中的重要特性。

定序资料，我们为方便起见可以将所有能按定序分类的

表 1.1　各社会阶层内的数目，约 1688 年

阶　　级	家庭数目
世俗贵族	160
精神贵族	26
从男爵	800
骑士	600
地主	3000
绅士	12000
官位较高的人	5000
官位较低的人	5000
从事海上贸易的著名商人	2000
从事海上贸易的一般商人	8000
法律界人士	10000
著名教士	2000
一般教士	8000

资料来源：格雷戈里·金，转引自 L. 索尔托夫："英国收入不平等的长期变化"，载《经济史评论》(L. Soltow, 'Long-run changes in British income inequality', Economic History Review)，第 21 卷(1968)，第 1 期，第 18 页。

资料以此称之，所以比定名资料更有价值，只是因为各类之间的次序是对资料的一种增加的信息，它以后可用于进一步的分析中。

1.3 区间或比率分类

正如各类的排列次序的附加信息使定序资料有别于定名资料那样，有关各类之间的精确关系的进一步信息是区间或比率分类的区别特性。当各类的次序和各类之间的区间规模已知，并且这一信息可用于以后的分析时，我们就能以这种方式将信息分类。对历史材料的计量分析中所使用的大部分资料属于区间或比率类型资料，最熟悉的例子包括收入资料、选举统计、投票人数、人口统计、作物收成等。如表1.2给出的就是英国1929年普选以后议会中各政党的情况。有了这种资料，不仅可以说议会中的工党议员比保守党议员多，而且可以说他们准确地多 28 名，而保守党则又比自由党多 201 名议员。换句话说，我们有一个固定的单位——议员的数目可用以测度各个政党的势力。假设测度的单位有一个零点，就像我们想象一个政党没有一名议员，那么我们得到的就是比率资料。只有当资料的测度单位不存在零点时，资料在理论上

表 1.2 下议院中各政党的势力，1929 年 5 月 31 日

政　党	席位数
工党	288
保守党	260
自由党	59
无党派人士	8

讲才属于区间资料;这种资料的主要例子是对温度的测度,其中零点是任意确定的。它不同于以货币单位测度收入的比率尺度,因为分文没有的情况具有不仅是任意确定的意味。不过,在实践中比率资料与区间资料之间的差别并不重要;大多数历史资料为比率类型,但它们常常也被看作是区间类型,在本书中我们可以交换使用这两个名词。

1.4 一些复杂的问题

经过初步分类之后,能够对资料进行进一步的分析的数量,依历史学家所掌握的资料的类型而定。区间资料比定名资料或定序资料更有价值,因为有关次序和排列的增加信息已蕴含在区间资料之中了,因此,对这类资料我们可以应用更复杂的分析方法。

正是由于这个原因,历史学家能准确判断他的资料是定名、定序、还是区间类型很为重要。如果他不能做到这一点,就会冒两种风险。如果他假定他的资料属定名类型而实际上是属定序或区间类型,这固然不成问题,但他却要为此付出代价,即他可能应用的分析技术的范围将受到极大的限制。另一方面,如果他假定他的资料属区间型而实际上却只是定序型,那么他将作出了一个错误的假定,他所应用的只适用于假定为区间资料的任何统计方法会产生错误的结果。因此,历史学家在开始他的分析之前,必须能够判断出他的资料所属的最近似类型。

在大多数情况下,能作出的或者我们能假定由资料的编纂者作出的分类类型是清楚的,像在本章前面所列举的定名、

定序和区间资料的例子那样。在另一些情况下，决定资料确属哪一种类型则要困难得多。举例来说，当格雷戈里·金作出根据社会阶层的英国人口一览表（如表1.1）时，他加上了每一阶层家庭年收入的估计值。表1.3给出了这些估计值的一部分，这次是为社会结构的下半部分所作。

至于表1.3中的左边两列，两者合起来显然代表定序资料。尽管对格雷戈里·金是否正确地确定了社会阶层的次序，比如陆军军官是否真属较海军军官为低的社会阶层，而两者是否又都低于店主这类问题可能存在着一些争议，但是这些资料属于定序类型是毫无疑问的。困难来自决定表中第三列（即以英镑为单位的每一家庭的年收入）是否属于一个不同的类型。从表面上看，它是一个比例尺度；英镑是一个明确的测度单位，具有一个清楚的零点。然而，问题在于格雷戈里·金编制他的估计值时所能得到的资料来源；由于他并没有关于1688年家庭收入的完整统计，他的估计值在很大程度上只能是基于自己经验的推测。我们必须决定金是否能够根据自己的经验作出准确的估计，或者他只是根据不同家庭在社会中的地位来定其可能的收入。如果属于前者，我们可以把他的资料看作是比例类型，尽管对确切的数字还有疑问；如果属于后者，那么只是假充比例资料的定序资料而已。

无法决定资料是属于哪一类型的问题在运用历史统计时发生得相当多，因为历史学家对他正在应用的资料是按什么方式编集的知之甚少。不幸的是，对处理这类问题并不存在着普遍适用的准则；历史学家必须自己作出判断，而他的读者又必须对这一判断作出判断。例如，应该注意到在上述金的资料这一特定事例中，年收入的次序并没有完全与社会阶层

的次序相一致；海军军官的年收入被说成是正排在他们前面的工匠和手工业者的一倍多，而年收入为 $42\frac{1}{2}$ 英镑的农场主又排在年收入为 60 英镑的从事文史哲和科学的人们之前。这种次序上的不一致或许可以证明，金还有某些其他的证据作为他的收入统计的依据，他并不是仅按照他对社会阶层的理解来定数字。因此，完全有理由将收入资料看作是比例类型的资料。

表1.3 英格兰按社会阶层分类的家庭数目和每一家庭的年收入，约1688年

阶　　级	家庭数	每一家庭年收入(英镑)
较高等的土地占有者	40000	91
较低等的土地占有者	120000	55
农场主	150000	$42\frac{1}{2}$
从事文史哲和科学的人	15000	60
店主和商人	50000	45
工匠和手工业者	60000	38
海军军官	5000	80
陆军军官	4000	60
普通士兵	35000	14
普通海员	50000	20
劳动人民和在户外做工的人	364000	15
茅舍农和贫民	400000	$6\frac{1}{2}$
流浪者，乞丐，吉普赛人小偷和妓女	30000人	每人 2 英镑

资料来源：G. 金。

一般说来，明智的是谨慎从事，假定资料是属于信息含量较少的类型，除非可以肯定它们符合信息含量较多的资料类型的准则。另一种办法是应用不止一种类型的统计方法而比

较其结果。本书后面还要举出关于类似的对象但适用不同资料类型的各种可能的方法的一些例子。

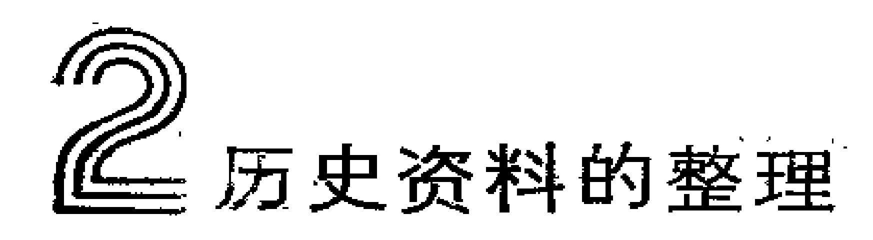
1.5　重新分类和编组

懂得同样一组信息可以按若干不同的方式进行分类是重要的。在我们所举的有关《末日裁判书》的例子中，我们得知威采邑有300头牧猪。就其本身来说，这是对信息的定名分类，但是如果我们掌握了，比如说，《末日裁判书》中其他采邑的牧猪数量的信息，我们就可以应用这一增加信息，根据其他采邑所拥有的牧猪数目，对威和其他采邑进行分类。如果我们只知道威拥有比另一村庄更多的牧猪，我们将得到一个定序分类，而如果我们确切地知道究竟多几头牧猪，我们就能够在区间分类里确定威的位置。这种分类和重新分类的能力非常有价值，因为只要我们时刻意识到已采用的分类类型，我们就可以用不同的方式利用资料的不同特征。

时刻意识到自己在做什么这一需要也适用于历史学家时常从事的另一种分类的形式。它不是一种替代，而是正常地与定名、定序和区间分类结合在一起的分类形式。在讨论定名分类时已经指出，如果有必要，资料的两个项目——庄园内的耕地和庄园外的耕地——可以归并为另一个类，即总耕地。这一归并过程，或有时称为综合，在历史学家应用和处理证据时时常发生。举一个最简单的例子，一个人既可以作为单个的人，也可以作为一个儿童组、父母组或祖父母组的成员之一。此外，根据他的工作，他还将成为一个职业组的一部分；根据他的年龄他又成为一代人的一部分，等等。如果我们充分了

解一个人,我们可以认为他或她处于某个一组或许多组之中。当我们把人们看作是城镇居民而这些城镇本身又是郡的一部分时，我们又可将上述各组归入一些更大的组。在这种情况下,我们既可以把他或她本人,也可以作为家庭的一个成员,城镇的一个居民或者国家的一个公民来谈论某个个人的行为。

在其他情况下,我们所知可能只限于一个组的行为,而对这个组里成员个人的行为却一无所知。那么我们掌握的是综合资料而不是个人资料。我们必须留意这一区别，特别是因为许多公布的证据是综合性的。例如，社会史学家大量利用人口普查的报告，在这些报告中描述了像特定教区或郡的居民这类集团的社会和经济特征。同样，经济史学家研究一消费者集团对某种商品的需求，而政治史学家则研究由投票者集团的行为所决定的选举结果。在所有这些事例中，集团的行为由此集团中所有成员的行为所决定,但是(除了出现未必可能的情形，即他们的行为完全一致)，我们无法推断出任何一个个人是如何行事的。换句话说，我们能够从个人资料导出综合资料，但是未必能从综合资料推断出个人资料。这在根据综合资料，试图将个人行为的某一方面与另一个人的行为的某一方面联系起来研究中是一个特殊的困难，所以在本书的后面还要再一次谈到这一问题。



2 历史资料的整理

历史学家除了应用上一章所讲述的方法将其资料分类之外，还必须学会整理这些资料以符合计量分析的要求。不同的统计方法需要不同的整理资料的方式，然而可以定出一些普遍性的原则和语汇。它们是为了保证清晰和一贯性，省时省力，并避免在分析的以后阶段中可能出现的混乱而设计的。

2.1 资料集

上一章中我们在一种最广泛的意义上用“资料”一词来描述历史学家所处理的材料。因而我们需要用另一个名词来表示那些用于某个特定分析项目的资料，为此我们将使用“资料集”这一名词来描述历史学家打算在某个特定分析项目中所使用的一组连贯的历史资料。如果一个历史学家感兴趣的问题不止一个，他可能想要在分析中使用几个资料集，但在这种情况下，他仍不妨把他的材料看成是一系列资料集，它们共同组成他正在研究的证据的整体。

把证据视为一系列资料集的目的，是要强调历史资料不应被看作是过去遗留给我们的一堆模糊不清的信息，而是与我们所想要研究的特定问题相关联的一件件信息。在任何一个研究项目中，我们都将选择我们认为与我们正在研究的问题有关的那些信息，而忽略其他信息。举些例子有助于澄清资料集的概念。如我们对1086年英国采邑社会模式的研究有兴趣，那么我们的资料集就很可能是对《末日裁判书》的考查；换句话说，我们从有关1086年英国的范围广泛的、各式各样的信息中选出一组信息，我们称它为这个特定项目的资料集。与此类似，在对英国选举的研究中，我们的资料集之一很可能是一组选举结果，就像表1.2列出的英国1929年选举结果一样。因而一个资料集就是从历史学家所能得到的所有历史资料中的一组连贯的资料。它之所以被选出是因为它与历史学家想要考虑的问题密切相关。

2.2 个案

每一个资料集都是由一系列个别的资料所组成，它们汇集起来形成一个与某个特定问题有关的证据整体。所以，在每一个资料集中，我们必须对资料加以整理使之便于对这个问题的思考；不能只把资料杂乱无章地写在纸上或卡片上，而必须加以一贯和合乎逻辑地整理。

对任何资料集进行整理的基本单位是"个案"。个案由与一项调查的特定部分有关的各件信息所组成。例如，在《末日裁判书》中，我们可以把每个采邑看作为一个个案，包含着描述这个采邑的各件信息。与此类似，在对选举结果的研究中，

每一次普选结果都可视为一个个案，因而1929年的选举结果就可以被认为是从各届英国普选结果资料集中所抽出的一个个案。因此，每个个案可能是一个人或一个采邑，也可能是选举结果，还可能是一些集团或集合体；通常，我们不将个人的和综合的资料混杂在任何一个资料集中。

2.3 变量

每一个案包括若干有关其自身的信息。这些信息描述了个案的不同特征。以《末日裁判书》为例，我们了解到在威采邑里有多少犁土地，多少名农奴，多少亩牧地。如果再看其他采邑，或其他个案，我们会发现有关其他采邑这些相同特征的信息；有些采邑可能拥有和威采邑同样多犁土地，其他采邑或多些，或者少些。"采邑拥有的犁数"这个特征将有变动，即从一个个案到另一个案有所不同，我们因此可以称它为一个可变化的特征，简称"变量"。我们因此可以知道每一个案都是由与所有个案共同的各种变量相关的若干不同的信息所组成，而且我们因此可以说，每个个案是由若干值所组成的，一个变量一个值。值并不一定是数量性的（比如每一采邑的名称，也是一个变量），而且采用数字和文字混合的方法来记录每个个案往往是合适的。

在我们普选结果的例子中，个案是普选结果，而变量为每次选举以后各党派的实力。因而有四种变量：工党实力、保守党实力、自由党实力，和无党派人士实力，而1929年这些变量的值分别为288、260、59、8。

2.4 资料矩阵

在纸上或者在脑中把资料集加以整理是合适的，这样做能使我们清楚地分辨出哪些信息是资料集的组成部分，哪些是个案，以及哪些是变量。使资料具有条理性的一个方便的方法就是使用“资料矩阵”。表 2.1 所给出的是《末日裁判书》采邑资料集的一部分，它就是一个资料矩阵的实例。

在表 2.1 中，我们就《末日裁判书》所给出的 5 个采邑提出了某些证据。我们提供出信息使得每一个案（在此表中就是每个采邑）都有自己的一行，而每一变量（每一采邑的可耕地数，草地的亩数，佃农的数目）都作为矩阵表的一列。所以，我们可以把资料矩阵认为是由几个通常代表个案的行和通常代表变量的列所构成的（在本书中，我们将严格遵守这一惯例，但是读者也应知道，为了展示方便或其他原因这一惯例有时会被打破）。

表2.1 《末日裁判书》采邑

采　　邑	可耕地	草地亩数	佃农
威(Wye)	52	—	114
斯迪夫基(Stiffkey)	1.5	2	—
密尔顿(Milton)	15	20	14
昂特尔(Oundle)	9	50	23
里兹(Leeds)	6	—	27

— 短划表示数据缺。

资料来源：J. J. 巴格利《历史的解释，1：英国中世纪史资料，1066—1540年》，哈芒兹沃斯：企鹅出版社，1965年，第 27—29 页。

我们可以把表 2.1 给出的矩阵称为具有 5 行 4 列的资料

矩阵。在分析过程中，将注意力集中于一行或一列，或者可能是某一信息上，而为了这部分的分析抛弃矩阵表中的其他内容，这对我们来说常是方便的，但我们这样做时，往往不知道该怎样称我们所感兴趣的那些特定信息；比如，说“给出昂特尔(Oundle)采邑中草地亩数的一个数据”就相当累赘了。

为便于指称资料矩阵中的单个信息，我们可以使用一种矩阵标记法。就像代数中我们常用字母表中的字表示数字一样，我们也可以用字表示矩阵表中的每一个信息(每一个“矩阵元素”)。表 2.2 现示了一种这样做的可能方法。

表 2.2 《末日裁判书》采邑

采　邑	可耕地	草地亩数	佃农
威(Wye)	a	b	c
斯迪夫基(Stiffkey)	d	e	f
密尔顿(Milton)	g	h	i
昂特尔(Oundle)	j	k	l
里兹(Leeds)	m	n	o

在此表中，a 代表52，k 代表50，与表 2.1 列出的矩阵表中的对应项目。利用这种方法，我们就能便于指称某个信息。现在我们可以用“k”代替“给出昂特尔采邑中草地亩数的一个信息”。

然而，用字母表中的字代表矩阵中的元素这种方法显然受到严格的限制。如果我们的元素超过 26 个，字母将会用光，而很多数据矩阵中的信息会超过 26 个。因此，我们需要某种更为概括性的方法来表示一个矩阵中的元素，表 2.3 显示的就是这样一种方法。

在此表中，我们用字母 A 并辅以下标表示矩阵中的每一

表 2.3　资料矩阵 A 代表《末日裁判书》5 个采邑

采　　邑	可耕地	草地亩数	佃农
威(Wye)	A_{11}	A_{12}	A_{13}
斯迪夫基(Stiffkey)	A_{21}	A_{22}	A_{23}
密尔顿(Milton)	A_{31}	A_{32}	A_{33}
昂特尔(Oundle)	A_{41}	A_{42}	A_{43}
里兹(Leeds)	A_{51}	A_{52}	A_{53}

元素，下标的第一位数表示元素所在的行，下标的第二位数表示元素所在的列。这种标记方法允许我们仅用字母表中的一个字母和两个下标来描述一整个矩阵，甚至一整个资料集和其中的每一个元素。我们可以选用不同的字母来描述不同的矩阵或资料集。

我们迄今已经讨论了不止 1 行和 1 列的矩阵，但是只有 1 行或仅有 1 列的矩阵也是可能的。例如，我们抽出表 2.1 中的第二行，就得到一个 1 行的矩阵，被称为 1 行向量，如表 2.4所示。

表 2.4　对斯迪夫基采邑观察的行向量

采　　邑	可耕地	草地亩数	佃农
斯迪夫基	1.5	2	—

此表可用表 2.5 的矩阵标记号代替之。

表 2.5　对斯迪夫基采邑观察的行向量 B

采邑	可耕地	草地亩数	佃农
斯迪夫基	B_1	B_2	B_3

注意我们用了一个不同的字母，以免与较大的资料矩阵 A 相

混淆，并且丢掉了一个下标。事实上，我们丢掉的是指出行的第一个下标；因为只有1行，第一个下标就是多余的了。

与此相类似，表2.1第一列也可以如表2.6用矩阵标记号表达，我们即得到了一列向量C。这里我们又用了一个不同的字母，这里丢掉了列下标；因为只有1列，第二个下标就是多余的了。

所以，同样的资料可被视为一个矩阵元素，或一个行向量元素或一个列向量元素。每次选择什么方式来表示它们完全取决于我们的兴趣在于整个矩阵，还是仅限于1个个案（行向量）或1个变量（列向量）。

表2.6 对《末日裁判书》5采邑可耕地数量观察的列向量

采　　邑	可耕地
威(Wye)	C_1
斯迪夫基(Stiffkey)	C_2
密尔顿(Milton)	C_3
昂德尔(Oundle)	C_4
里兹(Leeds)	C_5

一种更为有用的方法是用字母i代表行下标，用字母j代表列下标。据此，对于表2.3我们可以说，i下标的值为1,2,3,4或5，j下标的值为1，2或3。实际上，我们可将这个矩阵说成是矩阵A_{ij}，其中i的变化范围为1至5，而j为1至3。

我们将使用这类矩阵标记法来讨论本书后面描述的许多统计学方法。虽然乍看起来应用这种标记法似是引入不必要的复杂，但以后我们会明白应用这种标记法会大大简化对计量资料的处理。

2.5 收集资料

由于按数据矩阵形式整理的资料适用于按以后将要描述的方法进行分析，那么当然，除非有某些特别的考虑不用这类整理，我们应该以资料矩阵的形式收集和陈列资料以备分析。因此，历史学家在开始一项计量分析时，首先必须确定在他的证据中哪些部分将作为个案，其次确定哪些变量与他打算研究的个案相关。一旦确定了这两点，他就能按此整理他的材料。

用上述方法整理实际资料或多或少是一个复杂的过程，视基础资料的复杂性而定。把资料集整理成为资料矩阵的一个重要要求就是一致性；每一行必须由一个个案构成，而每一列记录必须包含与在这列中其他记录同属一类的信息。这种一致性一般容易达到；以表 1.3 中格列高里·金对以年收入划分的社会各集团家庭数目的估计为例，没有发生混乱的可

表 2.7 《末日裁判书》中4采邑需纳税土地

采　　邑	需纳税土地
威(Wye)	7苏伦(sulung)
密尔顿(Milton)	0.5海特(hide)
昂特尔(Oundle)	6海特(hides)
里兹(Leeds)	10卡勒凯特(carucates)及6波伐特(bovates)

8波伐特(bovates) = 1卡勒凯特(carucate)。①

① 苏伦(sulung，约合60—120英亩)，海特(hide，约合 120 英亩)，卡勒凯特(carucate，约合100英亩)和波伐特(bovate，1 卡勒凯特的 1/8)，皆为古时英国土地面积计量单位。——译者

能性。每一列都明确地规定，资料被清楚地说明。然而在其他事例中，或许由于原始资料记录上的混乱，也有可能导致错误。例如，《末日裁判书》收集的有关每一采邑最重要的信息之一就是对需纳税的土地而积的估计。表 2.7 所列出的是有关我们已经接触过的 4 个采邑的这方而的信息。

在此表中，资料指各采邑的同一特征，即每一个案中的同一变量。但是在记录资料时，如果我们简单地列为表 2.8 那样，那肯定要造成错误。

表 2.8 《末日裁判书》中4采邑需纳税土地

采　　邑	需纳税土地
威(Wye)	7
密尔顿(Milton)	0.5
昂特尔(Oundle)	6
里兹(Leeds)	10.75

在表 2.8 中，由于各个案记录土地面积的单位不同，我们的资料列是不一致的。凡此种种，都可能违反个案之间一致性的要求。例如，在对英国选举结果的研究中，若我们不仅以每次普选后各党派的情况，而还以每次补缺选举以后各党派的情况作为个案，那么我们得到的资料集就是不一致的。每一列中的信息也许是正确而一致的，但个案之间将不同；一个个案是普选竞争所有的议席以后的结果，另一个个案则是补缺选举（仅一个议席易手）的结果。

有些人也许认为收集资料并将它们整理成资料矩阵时强调一致性的要求是多余的，但当应用计量的分析方法时一致性则是必不可少的。当历史学家面临如表 2.7 所给出的那种不一致的资料集时，他在能开始分析资料前必须设法克服这

一困难。他面临着4种行动方式的选择。第一，也是最佳的选择，是将所有资料转换成一致的计量单位，如把苏伦和卡勒凯特转换成海特；不幸的是，由于可能不知道不同计量单位之间的关系，并非总能做到这一点。在经济史中，这类常见的问题之一即不同的布匹用不同的单位计量，而这些计量单位之间的换算率并非总为人知的困难。

如果不能将资料转换成一个共同标准，第二种可能性就是接受这些差异，而把资料记录在不同的列中，好似每一种面积的计量单位都是一个单独的变量，如表2.9所作。这种方法的困难在于难于进行跨个案的任何分析；而且它还浪费篇幅，特别是当需纳税土地只是需要记录的信息之一时，这一点可能是重要的。

表2.9 《末日裁判书》中4采邑需纳税土地

采邑	需纳税土地		
	(a) 苏伦	(b) 海特	(c) 卡勒凯特
威(Wye)	7	n.a	n.a
密尔顿(Milton)	n.a	0.5	n.a
昂特尔(Oundle)	n.a	6	n.a
里兹(Leeds)	n.a	n.a	10.75

n.a表示数据缺。

第三种可能性是省略非典型性的计量单位；苏伦为肯特郡的计量单位，不用于英格兰的其他地方，而卡勒凯特不如海特那样常用，因此省略以苏伦和卡勒凯特为单位的信息，而如表2.10那样记录信息或许是切合实际的。

由于抛弃了部分资料，这种方法比前两种方法较难令人满意；应用苏伦和卡勒凯特来计量资料毫无可能。

表 2.10 《末日裁判书》中4采邑需纳税土地

采　邑	以海特为单位记录的需纳税土地
威(Wye)	n.a
密尔顿(Milton)	0.5
昂特尔(Oundle)	6
里兹(Leeds)	n.a

n.a 表示资料缺。

第四种方法，也是最不令人满意的一种方法，是完全省略有关不一致记录的变量的任何信息。如在表 2.7 我们对《末日裁判书》研究的例子中，应用这种方法的结果就是失掉整个表格的意义——一种极端的办法。然而，若引起困难的项目只是需要记录的许多件信息之一，并且它在以后的分析中不占有很重要的地位，那么与其面临缺乏一致性的问题，完全抛弃它可能是无损的。反对这种方法的主要理由在于，资料在被抛弃，而在记录资料的阶段通常很难判断哪些信息在以后的分析阶段里是有还是没有价值。因此，除了在不得不这样做的情况下，保留资料总比抛弃资料为可取。

3 一些简单的数学方法

在这一章里，我们将讨论一些简单然而是重要的统计学方法的计算。我们还要讲述一些简单的数学技巧，它们简化了这些方法的计算，而且在本书的以后部分我们还需要用到它们。某些技巧对那些能记得他们在学校里学到的数学的人是熟悉的，而其他技巧则将对很多人来说是新的，对此我们将给予相当详细的说明。解释统计学概念也可以不用一两个这些数学技巧，但是这样做却可能导致不必要的复杂化和对简单的论点作冗长的解释。

3.1 频数分布

资料矩阵中的每一列都是由与本矩阵中各个案的某些变量特征相关的值所构成。倘若我们特别对某个变量特征感兴趣，那就集中注意力于矩阵中它所属的列，并在这一列中看到纵列数字，每一个数字对应一个个案。在上章《末日裁判书》的例子中，有一纵列关于需纳税的土地面积的数字，每个数字对应一个采邑。若如上章所示，我们考虑的仅是几个个案，那

么我们可以不很费力地了解有关每一采邑需纳税土地方面的信息。然而，如果我们在分析更多个案的资料，例如 50 个采邑，那我们就需从一个长得多的数字纵列（如表 3.1 所示）中了解和吸收信息。面对这样一张表，我们很难略为清楚地辨别出它所包含的信息的主要特征；我们不容易把 50 个或者更多数字所包含的信息吸收进我们的脑中。

因此为了吸收大量资料，我们需要以某种方式对资料加以概括，使它们的主要特征以一种便于记忆的方式示现出来。这样做的最简单方法，作为第一步，就是统计每一值在数字纵列上出现的次数。当这样做时，实质上我们是在重新整理资料；在原始的资料矩阵，如表 3.1 中的列向量，变量的值是以最初收集资料时个案在矩阵中出现的次序加以整理和排列的。而现在，我们改按个案所对应的变量值来整理它们。我们分布个案来显示各特定变量值出现的频数，因此对资料的这种重新整理被称为频数分布。

作为一个例子，我们来思考一下在对表 3.1 的资料进行分析时可能出现的问题。此表显示埃塞克斯郡 50 个采邑林地的牧猪数目，它们是 1086 年《末日裁判书》所记载的一部分。实质上，这些信息涉及的是这些采邑中的林地数量；《末日裁判书》的编纂者并未正规地给出林地面积数，而是计量了能够养活牧猪的头数。正如达比教授所说，“这些数字不一定反映在林地中牧猪的实际数目；猪在这里仅被作为一种计量单位。”① 知道了这点，我们可以见到表 3.1 为 1086 年埃塞克斯

① H. C. 达比：《<末日裁判书>时代东英格兰地理》(H. C. Darby, The Domesday Geography of Eastern England)，剑桥：1952年，第233页。

郡林地和森林面积提供了有用的证据。但如表现在所示，人们难以理解这信息。通过此表，我们无法了解到有关诸如每一地区林地分布的情况，以及是否存在某一块大家都不用的林地。为得到这类信息，我们需要对此表加以概括，把它所包含的信息变成一张易于处理的表格。

表 3.1　埃塞克斯郡林地，1086 年

地　　　　区	1086年代表林地面积的猪数目
里特尔(Writtle)	1200
克拉弗林(Clavering)	600
法恩哈姆(Farnham)	150
〃　〃	50
乌格莱(Ugley)	160
阿尔费勒斯图那(Alferestuna)	350
肯菲尔特(Canfield)	120
邓莫(Dunmow)	300
伊斯顿(Easton)	150
〃　〃	400
〃　〃	150
拉希莱(Lashley)	60
撒克斯蒂(Thaxted)	800
亚特莱(Yardley)	30
赫沙姆(Hersham)	30
哈林伯里(Hallingbury)	100
芬钦菲尔特(Finchingfield)	5
〃　〃	30
海廷哈姆(Hedingham)	500
〃　〃	160
海尼(Henny)	30
〃　〃	20
马布尔斯泰(Maplestead)	60
〃　〃	15(原 16)

续表

地　　区	1086年代表林地面积的猪数目
波尔海(Polhey)	40
萨林(Saline)	200
斯丹斯特(Stansted)	400
韦瑟斯菲尔特(Wethersfield)	500
威克姆圣保罗(Wickham St. Pauls)	20
伊斯特伍德(Eastwood)	30
安伯登(Amberden)	200
伯钱杰尔(Birchanger)	50
” ”	30
埃尔森哈姆(Elsenham)	1000
萨弗隆沃尔登(Saffron Walden)	800
” ”	30
塔克莱(Takeley)	600
桑特莱(Thunderley)	600
威克姆蓬亨特(Wickham Bonhunt)	80
温比希(Wimbish)	60
莱厄(Layer)	400
” ”	60
科格沙(Coggeshall)	30
布拉克斯蒂(Braxted)	500
诺特莱(Notley)	80
” ”	30
” ”	200
” ”	100
里文哈尔(Rivenhall)	350
未特别指明的拜斯特勃尔亨特赖特(Barstable Hundred)	55
	11915

资料来源：H. C. 达比：《<末日裁判书>时代东英格兰地理》(H. C. Darby, The Domesday Geography of Eastern England)，剑桥：1952年，第236—237页。

我们可以通过建立一个像表 3.2 那样的简单的频数分布来开始对表 3.1 进行概括。现在，我们有 44 个数字，而不是表 3.1 中一个数字对应一个地区的 50 个数字，其中第 1 列中的 22 个数字表示牧猪数变量的值，第 2 列的 22 个数字表示这些值在资料中出现的频数。以表 3.2 与表 3.1 相比较，虽则我们对需要吸收的信息量有所减少，但并没有简化多少。人们仍难于一看就理解它的内容。

表3.2　埃塞克斯郡各教区根据牧猪头数的频数分布，约1086年

1 牧猪头数	2 拥有第一列牧猪头数的地区数
5	1
15	1
20	2
30	9
40	1
50	2
55	1
60	4
80	2
100	2
120	1
150	3
160	2
200	3
300	1
350	2
400	3
500	3
600	3
800	2
1000	1
1200	1

然而，我们可以建立其他类型的频数分布来进一步帮助我们理解这类资料。例如先将变量值编组，再将频数排列起来，并将所属个案归入每一组；其结果就被称为一种编组频数分布，例子如表 3.3 和表 3.4 所示。在很大程度上我们可以随意选择不同的值的编组；若我们要强调个案之间的细微差别，可以多编几个组；如我们感兴趣的只是大的差别，则可以用少量的编组。各组的大小并不一定要一致，但不一致可能引起混乱；除非有某种不可克服的困难，应该用大小一致的编组。对各组唯一的绝对要求是它们必须没有歧义，即不允许发生一个个案应属哪组的争执。因而各组永远应像表 3.3 中那样说明：0—199，200—399 等；如果它们作 0—200，200—400 等那样的说明，则不知应把一个拥有 200 头牧猪的个案

表3.3　埃塞克斯郡各教区根据牧猪头数的编组频数分布，约1086年

牧猪数	地区数
0-199	31
200-399	6
400-599	6
600-799	3
800-999	2
1000-1199	1
1200-1399	1

放在哪组，因而引起混乱。

频数分布是一个重要的统计学工具，我们将在下一章里回过来对其应用加以较详尽的讨论。现在只需记住，频数分布是对资料的一种再整理——个案按照它所具有的变量像而排列。

表3.4 埃塞克斯郡各教区根据牧猪头数编组频数分布，约1086年

牧猪数	地区数目
0-99	23
100-199	8
200-299	3
300-399	3
400-499	3
500-599	3
600-699	3
700-799	0
800-899	2
900-999	0
1000-1099	1
1100-1199	0
1200-1299	1

3.2 求和记法

在对计量资料的分析中，我们常想要计算和应用一个数字集的总数。将数字相加得到一个总数，正规地称为"和"的过程是大家熟悉的，但是，必须写出如"计算变量值之和"这种指示往往显得笨拙。因此，用一个记号作为求和的指示是有用的，它也可以在以后的计算中被用来代表总和。

在求和记法中，希腊大写字母西格马($\sum$)表示数字相加，而字母上下和右边的其他项说明什么与什么相加以求和。以表3.1为例，如我们把牧猪数目的纵行视为一个列向量，并用字母X来表示，其值X_i从X_1直到X_{50}，那么我们可以把这一列之和表示为

$$\sum_{i=1}^{50} X_i = 11915$$

$\sum$ 下面的项 $i=1$，表明X向量的下标 i 从 1 开始，取值并随后取连续的正值 2、3、4、5，等等。$\sum$ 记号上面的项表示 i 的最后取值，在本例中为 50，因为表中有 50 个个案。$\sum$ 右边的项 X_i，表示要求和的向量。如果我们愿意，可以改变 $\sum$ 上面和下面的项，以表明仅对纵列的一部分求和。比如，我们打算从计算中剔除里特尔，克拉弗林，法恩哈姆，里文哈尔，以及拜斯特勃尔亨特赖特中"未特别指明的"采邑，就可以写出下面的式子来表达对其他采邑的求和：

$$\sum_{i=5}^{48} X_i$$

它指示我们 i 从 5 开始取值，随后连续取正值至 48 求和，从而剔除了纵列中最前面的 4 个和最后面的 2 个个案。

对包含任何个案数目的一个纵列，指示的一种有用的普通形式是把其个案数目称为 N，并把求和指示写为：

$$\sum_{i=1}^{N} X_i$$

有时我们想要不仅得到一个向量的总数，面且还要得到一整个矩阵的总和。当我们按两种方式对资料集进行分类，面

表 3.5　1851 年中期英伦三岛人口估计　　单位：千人

	男性	女性
英格兰和威尔士	8,809	9,174
苏格兰	1,379	1,517
爱尔兰	3,181	3,333

资料来源：B. R. 米切尔和 P. 迪恩：«英国历史统计摘要»(B. R. Mitchell and P. Deane, Abstract of British Historical Statistics)，剑桥：1962 年，第 8 页。

构成像表 3.5 那样的资料矩阵时，矩阵总和往往是必不可少的。为得到 1851 年英伦三岛人口总数，我们需要对矩阵的所有元素求和。

如果我们将表 3.5 中的资料用矩阵Y来表示，就得到一个具有在表 3.6 中所显示的元素的矩阵。

表3.6　表示1851年中期英伦三岛人口估计的材料矩阵 Y

	男性	女性
英格兰和威尔士	Y_{11}	Y_{12}
苏格兰	Y_{21}	Y_{22}
爱尔兰	Y_{31}	Y_{32}

我们可将矩阵Y中所有元素的和写成

$$\sum_{i=1}^{3}\sum_{j=1}^{2}Y_{ij}$$

这里在具有 i 行和 j 列的矩阵Y中，第一个$\sum$指行，第二个$\sum$指列。全部指示因而是对矩阵中所有元素求和。我们将用的把元素相加的实际步骤将从 $i=1$ 和 $j=1$ 开始，若矩阵以一个表来表示，则首先取左上角 Y_{11} 为元素。然后取 $i=1$ 和 $j=2$ 的第二个元素 Y_{12} 相加，接着再转向第二行取元素 Y_{21}。换句话说，第二个$\sum$的下标在第一个$\sum$下标的每一个连续值范围内变化。

求和记法的目的在于简化计算，及简化对统计学公式中数字之和的使用。在把 X 定义为表 3.1 中所示的列向量后，我们不写“在表 3.1 中所列数目的总和”，而可以只写成

$$\sum_{i=1}^{50}X_i$$

实际上，求和记法也常通过省略下标而进一步简化。比如，当我们想表示向量Z之和时，只写$\sum z$，而不是

$$\sum_{i=1}^{N} Z_i$$

但重要的是，这样做不应引起混乱，而一般还应使用下标。

应该指出，当$\sum$后带有由括号括起的数字符号组成的项时，它表示该项所包含的是所求之和，直到碰到一个+号或－号为止。例如，对两个等价向量X和Y的元素相乘，然后对其结果求和，可写成：

$$\sum_{i=1}^{N} X_i Y_i$$

它表示为从1至N的每一个i值，我们把X_i乘以Y_i并求和。但如果我们想把其他数加到这个和上，比如55这个数，可写成

$$\sum_{i=1}^{N} X_i Y_i + 55$$

这显示我们要把55加到X与Y向量的乘积之和上，而不是把55加到每一个X_i与Y_i的乘积上，然后求和；如我们要做的是后者，应写成

$$\sum_{i=1}^{N} (X_i Y_i + 55)$$

事实上，+号表示求和应停止在什么地方。

我们可以像使用任何其他代数量那样使用求和记号及与它相关的符号。这样我们可写出：

$$K \sum_{i=1}^{N} X_i Y_i$$

表示计算完X_i与Y_i乘积之和以后，我们用另一个量K与那个和相乘。在统计工作中常遇到的应用求和记法的例子是

$$\sum_{i=1}^{N} X_i^2 = (X_1^2 + X_2^2 + X_3^2 + X_4^2 \cdots\cdots + X_{N-1}^2 + X_N^2)$$

及

$$\left(\sum_{i=1}^{N} X_i\right)^2 = (X_1 + X_2 + X_3 + X_4 \cdots\cdots + X_{N-1} + X_N)^2$$

在这些例子中，一排点表示应连续对点之前(在上述例子中为 X_4^2 和 X_4)直至数列终点(X_N^2 和 X_N)之间的所有值求和。

3.3 对数

历史学家常须计算和分析比例变化；例如，我们可能想要查考英国一个年度与第二个年度的出口变化，把它作为前一年的比例或百分比。或某一政党从一次选举到下一次选举得票变化的比率。许多诸如此类在第六章中讨论的问题，都适宜于利用对数。对数的概念对大多数学习过初等数学的读者来说是熟悉的，虽则在电子计算器时代简化多位数运算这一对数的主要用途显得不那么重要了。然而，在统计学中对数仍十分重要，因此我们有必要先回顾一下对数用法的主要特点。

在统计学中使用两种类型的对数，一种是以10为底的对数，另一种是以 e 为底的对数。本书只涉及前者，因为它更为常用。我们可以利用本书最后附录里的四位数对数表来查找一个数的常用对数。当使用这些对数表时，我们仅取数的前四位有效数字(即第一个数不为零)。对于多位数字，此表则有失精确，使用电子计算器可以解决这个问题，许多类型的电子计算器只要一揿电钮就能计算对数。当然使用四位数对数表依然是明智的，因为对数的基本概念及其使用方法在此显得更为清楚。

例如，查找104,869.0的对数，对第5位数四舍五入以后，取前四位数就得到1049。此数的前两位要我们去查对数表的行，在本例中为行10。第3位数为4，沿着行10查到有以4为首的四位数字的那一列；第四位数为9，仍沿此行查到有以9为首的一位或两位数字的那一列。在第四列中我们得到.0170，在标有9的列中得到值.0037，两者相加为.0207。同理，若我们想计算1272的对数，可以从对数表的第3行得到结果：.1045。

然而，在查找上述值时，我们仅有这些数字的对数的一部分(称为尾数)。此外，我们还要考虑到小数点的位置。为此，我们必须了解(如“以10为底”这个名称所隐含着的)10的对数是1.000，100(10^2)的常用对数是2.000，1000(10^3)的对数为3.000，以此类推。因此一个介于10和100之间的数其对数值必在1.0000和2.0000之间。在这两个界限之间的精确值由尾数给出，如我们已显示过的，查对数表而得。对数中决定小数点位置的那部分称为首数，以及它与小数点无关的那部分对数，即尾数，一起组成了我们计算中所应用的对数。表3.7显示小数点位置的变化所带来的影响。

表3.7　小数点位置对对数的影响

$\log 1272.0=3.1045$	$\log 0.12720=\bar{1}.1045$
$\log 127.20=2.1045$	$\log 0.01272=\bar{2}.1045$
$\log 12.720=1.1045$	$\log 0.001272=\bar{3}.1045$
$\log 1.2720=0.1045$	$\log 0.0001272=\bar{4}.1045$

考虑小数点位置(从而决定对数的首数)的最简便的方法为：考虑在真数得到一个有效数字之前小数点需向左边移动的位数。如在表3.7中，我们必须把真数127.2的小数点向

左移两位才得到 1.272，因此首数就是 2。与此相类似，对于小于 1 的真数，我们要计算需将小数点向右移动的位数。比如，对于 0.0001272，必须将小数点移四位，因此首数为 $\bar{4}$。

确定了我们要用的数的对数值以后，就可以实施我们所要做的数学运算了；对此我们即将加以阐述。运用对数进行运算所得的答案本身就是一个对数，因此必须通过使用反对数表(见本书后面所附)将其转换就使我们得到最后的结果。假设我们的对数值为 2.7127。我们取尾数，在反对数表中查到标有 0.71 的那一行；然后在有四位数的第 3 列中找到 5152；再从有一位或两位数的第 7 列找到 8。5152 + 8 = 5160。现在我们得到一个四位数的真数，利用前边得到的有关对数首数的知识，我们还必须确定小数点的位置，在本例中首数是 2。它告诉我们应从小数点左边只有一个有效数字的位置开始把小数点向右移动两位(在前面的对数中，需将小数点向左移动两位才能得到首数 2)。因此，对于 5160，小数点位置应从 5.160 开始，然后向右移两位，得到 516.0。2.7127 的反对数为516.0。

能用对数进行的运算最容易排列成表，如表 3.8 所示。

应注意，对数首数上边所置的一小横，如 $\bar{2}$，仅表示对数是一个绝对值(没有加减符号)小于 1 的数。在对对数进行运算时，我们省略了正负号，仅在运算的最后才加以考虑。如 274.6 乘以 − 58.27，实际运算如表 3.8，直到计算结束才将负号放回所得出的答案 − 16,000.0。

本章前三节已讨论为了理解本书后面的内容，必须掌握的所有数学运算方法。实质上，想要采用计量方法的历史学家必须知道如何加减乘除，计算平方和平方根，以及应用简单

表3.8 对数运算

运　　算	方　　法	例　　　　子
274.6×58.27	对数相加,求反对数	274.6×58.27=log(274.6) +log(58.27)=2.4387 +1.7654=4.2041 (求反对数)=16000.0
274.6÷58.27	对数相减,求反对数	274.6/58.27=log(274.6) −log(58.27)=2.4387 −1.7654=0.6733 (求反对数)=4.713
274.6的平方	对数乘2,求反对数	247.6^2=log(274.6)×2 =2.4387×2 =4.8774 (求反对数)=75410.0
求58.27的平方根	对数除2,求反对数	$\sqrt{58.27}$=log(58.27)/2 =1.7654/2=0.8827 (求反对数)=7.633
求0.9854的平方根(小于1的数)	对数除2,求反对数	$\sqrt{0.9854}$=log(0.9854)/2 =$\bar{1}$.9936/2 =($\bar{2}$+1.9936)/2 =$\bar{1}$+0.9968=$\bar{1}$.9968 求反对数=0.9926

的矩阵和求和记法。为简化或加速运算，可以使用第九章所谈到的计算器或计算机之一；它们可以消除很多与手工计算有关的困难,但即使应用它们,历史学家仍然必须懂得本章所讲的那些简单的数学运算。一般来说，对方法的选择取决于必须实施的运算类型，以及所要求的精确度。如果需要高精确度,必须应用手工方法或能够处理大量数据的机械方法;若经过四舍五入后的四位有效数就能满足要求，那么就可以使用对数和平方根表。

4 资料的初步分析 I：频数分布法和图表

按照前几章所讲的方法收集和整理资料后，研究项目的分析阶段就可以开始。计量分析的最初阶段因项目和历史学家的不同而异，但是不妨说，在初期阶段历史学家需要应用"描述性统计"方法。描述性统计是那些主要涉及资料的组织和描述的统计方法；这种统计方法有时常与"分析性"统计相对比，但这样的区分是不确切的，本书不打算使用它。我们将讨论的描述性统计学既是对资料分析的一部分，也是更高级统计方法的一部分。

描述性统计的作用是便于对计量材料的理解。它们可以帮助历史学家继续他的分析，也可以帮助读者理解分析的结果，但两者的目的都是为了获得更多的理解。既然如此，描述性统计应根据它们成功地增强理解的程度来加以评价；所以，虽然本章下面讨论的几种应用描述性统计的方法可能使粗心大意者发生误解，但是就应用这些方法而论却无所谓正确与错误。的确有某些只适合于某些特定类型的资料的描述性统计方法；例如，定序和定名资料就不能计算平均数。但除了这些例外，我们应只着眼于哪些方法能最清晰地阐明我们最感

兴趣的那些资料的特征来选择表述我们的材料的方法，也即我们所用的描述性统计。

应该强调指出，描述性统计方法不仅在呈现结果时有用，而是在分析的每一阶段都有用的。画一张曲线图，只需几分钟的工夫，或许能一下子显示出资料的某些方面，而当历史学家只是对着一张数字的表格看时是根本看不清的。

表 4.1 所示是资料被收集后未经任何重新整现或统计学上的处理的一个资料矩阵的某一典型部分。表 4.1 中的资料集由 1907 年英国商船队中约 25 艘船的有关数据组成，每一只船都标有“官方号码”(一个独特的识别号码，就像每辆汽车都有自己的号码牌照一样)。表 4.1 按列的次序包括每一个案的识别号码，两个定名变量，和两个区间变量。表内没有任何定序资料；在历史研究中极少遇见定序资料，因此没有必要详细地考虑它们的范例。每当一种特定类型的统计方法适用于定序资料时我们会提到这一情况；有关定序资料更为详细的讨论请参阅为社会科学家编写的统计学教材。

4.1 频数分布

在第三章里，我们介绍了描述性统计中应用最为广泛的方法，即频数分布。我们说明了频数分布基本上是将资料矩阵或矩阵中的个别列重新整理成一种使它所包含的信息更容易被理解的形式。重新整理可以仅涉及以一种新的方式安排个案，也可以按照个案的各种变量特征的值进行编组。频数分布可从定名资料、定序资料，或区间资料中建立。作为例子，我们将应用有关动力方式的资料——一种定名变量，及有

表4.1　25艘英国商船，1907年

官方号码	贸易对象	动力	吨位	船员人数
1697	国内	未知	44	3
2640	国内	未知	144	6
35052	国内	未知	150	5
62595	国内	风帆	236	8
73742	国内	蒸汽	739	16
86658	国内	蒸汽	970	15
92929	国际	蒸汽	2371	23
93086	国内	蒸汽	309	5
94546	国际	蒸汽	679	13
95757	国内	风帆	26	4
96414	国际	蒸汽	1272	19
99437	国际	蒸汽	3246	33
99495	国内	蒸汽	1904	19
107004	国内	蒸汽	357	10
109597	国内	蒸汽	1080	16
113406	国内	蒸汽	1027	22
113685	国内	未知	45	2
113689	国内	未知	62	3
114424	国内	风帆	68	2
114433	国际	蒸汽	2507	22
115143	国际	风帆	138	2
115149	国内	蒸汽	502	18
115357	国内	蒸汽	1501	21
118852	国际	蒸汽	2750	24
123375	国内	蒸汽	192	9

资料来源：根据船舶和海员登记总署的船员清单。

关船员数目的变量——一种区间变量。

资料矩阵中的第3列动力方式，是一个定名变量，可以取蒸汽，风帆和未知三种值之一。所以，根据此变量建立一个频数分布，只需计算在资料矩阵第3列中每一类型的动力出现

的次数，然后将结果组成一张新表，如表 4.2。

表4.2　表4.1第 3 列资料的频数分布

动力	船数
风帆	4
蒸汽	16
未知	5
总计	25

请注意，我们在表标题上说明本表所含资料的来源。为有助于准确性及理解，我们还给出了表中个案的总数。

相对于动力方式的三种可能值，区间资料变量，船员数（表 4.1 中第 5 列），从理论讲有很多的可能值。那个时期某些商船上船员人数达几百人之多，可是在我们的案例中船员数目仅为 2—33，因此可将我们的频数分布限定在这些值内。即使有了这个限定，若我们模仿表 4.2 的方法就会得到一个有 32 个可能值的表，其中大多数在第 2 列中的值为零。因此为简化和压缩频数分布，我们将值进行编组并计算出属于每一组的个案数。结果如表 4.3。

又要注意到，表 4.3 第 1 列中各编组规定得不可能产生一个个案究应属于一个编组还是属于另一个编组的混乱；如各组按 0—5，5—10，10—15 等等规定，就会发生混乱。

表 4.3 的建立立即使我们对资料有了一个比我们会从表 4.1 的一堆数字中所能获得的清晰得多的概念。如果愿意，我们还可以建立其他类型的频数分布，用以阐明资料的某些个别的特征。百分比频数分布是一种常用的频数分布，适合于所有类型的资料。在这类表中，频数不用绝对数来表示，即

表4.3　根据表4.1第5列中资料的编组频数分布

船员人数	船数
0-4	6
5-9	5
10-14	2
15-19	6
20-24	5
25-29	0
30-34	1
共计	25

每一频数出现的次数，而是按所占个案总数的百分比来表示。表 4.4 和表 4.5 就是这类百分比频数分布。在每种情况里，表中项目的总数都是 100。

表4.4　根据表4.1第3列资料的百分比频数分布

动力	占商船总数的百分比
风帆	16
蒸汽	64
未知	20
总计	100*

* 个案总数为25。

在建立百分比频数分布时，需要适当注意。百分比本身就是一种描述性统计，如果个案总数很小，百分比会给人一种错误印象；如果情况是这样，那么频数的很小差别可以由于转换成百分比而被扩大。因此永应给出个案的总数，如表 4.4 和表 4.5，使读者能将百分比转换成绝对数，如果他要这样做的话。

表4.5　根据表4.1第5列资料的百分比编组频数分布

船员人数	占商船总数的百分比
0-4	24
5-9	20
10-14	8
15-19	24
20-24	20
25-29	0
30-34	4
总计	100*

* 个案总数为25。

累积频数分布和累积百分比频数分布，是有时有用的频数分布亚型，虽则它们只适用于定序和区间资料。当需要了解有多少个案是在某些特定值之上或在其下时，上述方法是有益的。例如表4.6和表4.7所示。

表4.6　根据表4.1第5列资料的累积编组频数分布

船员人数	商船数目
4 或不到	6
9 或不到	11
14或不到	13
19或不到	19
24或不到	24
29或不到	24
34或不到	25

注意在表4.6和表4.7里没有给出总数，因为不言而喻频数列中最后一个数字必然是个案总数，亦即在百分比分布中为100。

表4.7　根据表4.1第5列资料的累积编组百分比频数分布

船员人数	占商船总数的百分比
4或不到	24
9或不到	44
14或不到	52
19或不到	76
24或不到	96
29或不到	96
34或不到	100*

* 个案总数为25。

4.2　交叉分类

迄今为止我们讨论了在资料矩阵中的这列或那列里运用频数分布对资料进行概括。我们还可以使用类似于建立频数分布的方法，在不止一列里对资料进行概括。其结果就称为交叉分类。

表 4.8 是最简单的交叉分类形式，其中一个定名变量(动力)对照一个区间变量(吨位)进行分类。像表 4.8 这样的表有时也被称为根据第 3 列和第 4 列所制的表。

对表 4.1 中的所有变量都可进行类似的交叉分类，依照

表4.8　根据表4.1第4列及第3列资料的交叉编组

动力	吨位
风帆(4艘)	468
蒸汽(16艘)	21406
未知(5艘)	445
总计(25艘)	22319

所涉及的资料类型和每一变量的可能值的数量，将其结果置于不同的表格形式之中。表 4.9 被称为列联表，表中的记录表示对应小标题下所示值的个案出现的次数。因而列联表可以被看成是按两种或更多方式进行分类的频数分布。

表4.9　根据表4.1第2列和第3列的列联表

动力	贸易对象		总计
	国内	国际	
风帆	3	1	4
蒸汽	10	6	16
未知	5	0	5
总计	18	7	25

正如我们可以有百分比频数分布一样，我们也可以建立百分比列联表，如表 4.10。在表 4.10 中，表 4.9 的每一个记录都以商船总数的百分比来表达。如我们想要阐明资料的某些特征，也可以建立百分比列联表，其中的记录不是按照总计（在本例中为 25 ）的百分比，而是按照某些编组的个案总数的百分比来计算。表 4.11 就是这一种方法的例子。

表4.10　根据表4.1第2列和第3列的百分比列联表

动力	贸易对象		总计
	国内	国际	
风帆	12	4	16
蒸汽	40	24	64
未知	20	0	20
总计	72	28	100*

* 个案总数为25。

在表 4.11 中，表中的记录以列总数的百分比来表示；而并未给出行总数，因为它们毫无意义。作为另一种办法，也可以计算行总数的百分比，在这种情况下则不给出列总数。注意，相除和四舍五入的过程中在第 1 列造成了一个稍微超过 100 的总数。

也可以建立对 3 种或更多变量加以概括的表，但这样做的危险在于表会像原始资料那样模糊不清和难以解释。除非有某种特别的理由，汇总表应限制在仅有 1 个或 2 个变量的资料上。

表4.11　根据表4.1第2列和第3列的百分比列联表

动力	贸易对象	
	国内	国际
风帆	16.7	14.3
蒸汽	55.6	85.7
未知	27.8	0.0
总计	100.01	100.00*

* 个案总数为25。

我们介绍了表 4.2—表4.11 而并没有对一种制表方法相对于另一种制表方法的优点，或对每一种方法所揭示的资料的具体特点加以评论。然而，应该清楚的是，每种方法都揭示资料的不同侧面；如表4.4表明将近三分之二的商船以蒸汽为动力，表4.7表明在一半以上的商船上船员人数不超过15人，表4.9则告诉我们从事对外贸易的商船中仅有 1 艘为风帆船。资料的所有这些侧面对一个致力于商船史的研究者来说都是有价值的；仅靠查看表4.1中未加处理的资料矩阵，没有一个侧面会立即显露出来。因此，实际中选择采用哪种制表方法

取决于人们探讨资料的哪一侧面（而在某种程度上也取决于资料是否按定名或区间标准分类）。

4.3 图表

在这一章里我们集中于运用列表方法来呈现资料的讨论。存在着许多其他呈现资料的方法；特别是我们可以用某些形式的图表来呈现资料。许多人发现如果把证据以某种方式的图表绘制出来。那么它的多重含义就更容易为人所了解。因此，描述性统计学的图表方法在呈现分析的最终结果上特别有用。而且，处理计量材料的历史学家可以从在分析的预备阶段以图表方式表达他的结果中得到好处。以这种形式表达的资料可能显示出历史学家始所未料的类型，而这可能导致进一步分析的设想。

呈现定名资料（如果要的话，也可以用于定序和区间资料）最普通的方法是运用条形图，如图 4.1 所示。在此图中，我们用商船动力方式表示的航运例子的资料以条形图的形式来予以呈现。称为条形图中的条子彼此完全分离，以强调我们在绘制的是定名资料，各类之间没有定序或区间关系这一事实。由于所绘制的是定名资料，沿水平轴的条子的次序无关重要；即使把它们打乱，也不会损失或改变所要呈现的信息。另一方面，如果所绘制的是定序资料，虽非绝对必要，但按常规要把各条形按类或变量的次序沿水平轴排列。并请注意，个案据以分类的变量（图 4.1 中为动力方式）应沿条形图的水平轴标绘。在条形图中，沿垂直轴显示分类变量，而沿水平轴显示条形并非错误，但通常条形是沿垂直轴显示的，如图

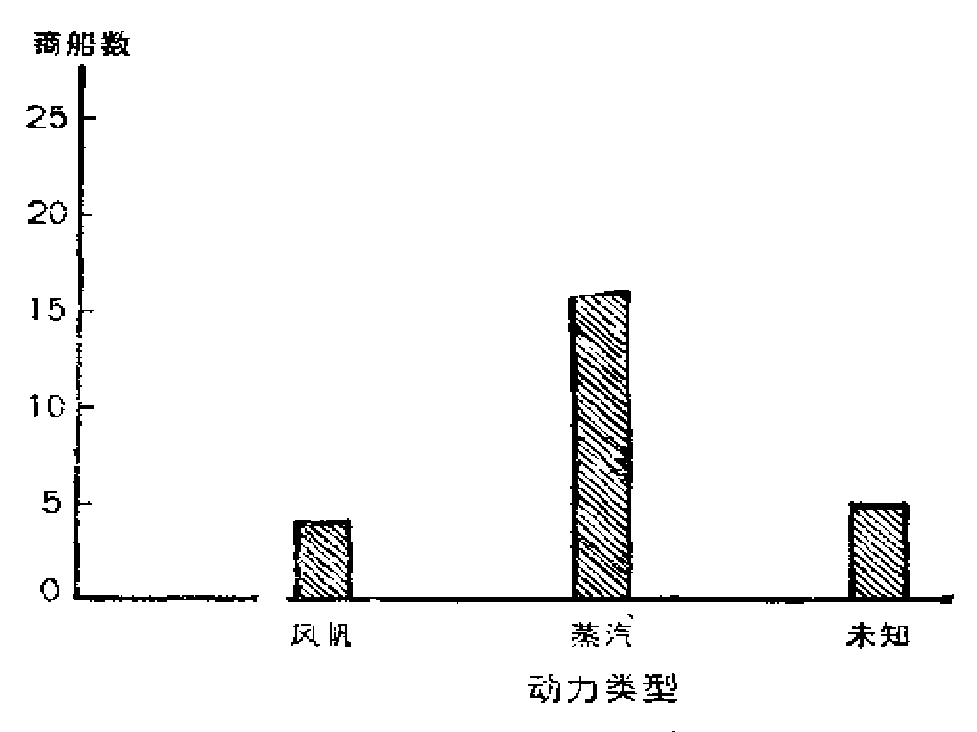


图 4.1　根据表 4.1 第 3 列资料的条形图

4.1所示。

如果要绘制的是区间资料，那么我们也可以用条形图的形式来呈现之，但用直方图则更为正规，如图 4.2。由于是区间资料，资料值就不像图 4.1 中那样彼此分开，而是显示为沿水平轴依次相接。在本例中，分类变量(船员人数)的不同值不仅像在条形图中那样以条的高度来显示，而且以柱的面积来显示；因此，重要的是直方图中每一柱的宽度应保持一致，以使面积与所呈现的频数成比例。否则会给人以错误的印象。

区间资料也可以用图解形式来呈现。图 4.3 表明怎样把表 4.1 中关于船员人数的资料用图表示。我们取船员人数的各个编组，即资料据以分类的变量，并把这些沿水平轴显示；各船员人数编组出现的频数沿垂直轴显示。这种图实质上是

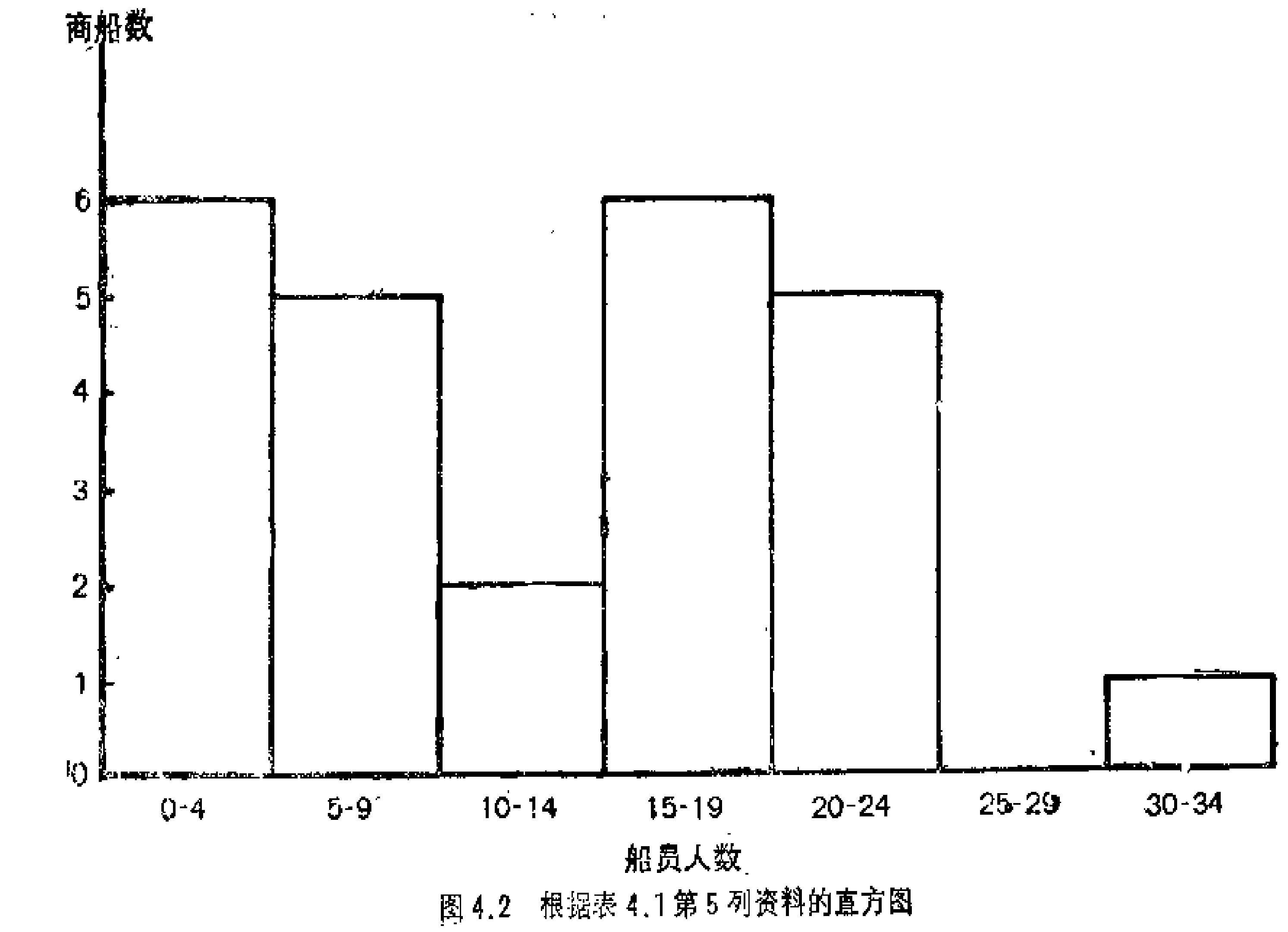


图 4.2　根据表 4.1 第 5 列资料的直方图

一个简单的编组或未编组的频数分布的图解表格。

当区间资料被制成图时，也通常对图表的形式略加改动，将图上各点连结起来，造成有时所称的一个曲线图。理论上，我们只应把图中的各点连结起来，从而给人一种变量具有连续性的印象——如果我们准备作这样的假设，即变量实际上是连续的，意谓在理论上变量可以取任何值。这类假设常可在涉及科学资料的案例中作出，例如温度和距离都属于这一类型的变量。另一方面，大多数历史资料是非连续性而是分立性的；变量的可能值呈阶段性变化。例如，若我们衡量人口总数，就必须按 1 个人的倍数来衡量；而不能像对气温的度数或距离的公里数那样细分。另外，很多历史资料虽然在理论上是连续的，然而由于计量的不精确在实践中却并非如此。这方面的例子如人的年龄；我们很少知道我们所研究人物的精确到天数的年龄。尽管从理论上讲如果有更多的信息我们可以做到这一点。

由于所描述的大多数历史资料无论从理论上还是在实践中都是非连续性的，似乎我们不应正规地利用曲线图。然而，事实上由于运用曲线图所获得的理解是如此之重要，以致我们能感到完全有理由应用它们，常记住我们必不能把资料看成具有连续性。作为一个例子，我们可以研究图4.4，一个根据图4.2制成的线图。先看图中标有A的点，它位于水平轴上标有5-9那一点之上，亦处于相对于垂直轴上标有 5 的水平线上。显然A表示有 5 艘船的船员人数分别为 5、6、7、8 或 9 。图中B点处于与A点相同的水平线上，但是它处于介乎水平轴上标有10-14和15-19的两点之间的垂直线上。因此它不代表放在图下面的船员人数的频数分布编组中的任何一组。我们不

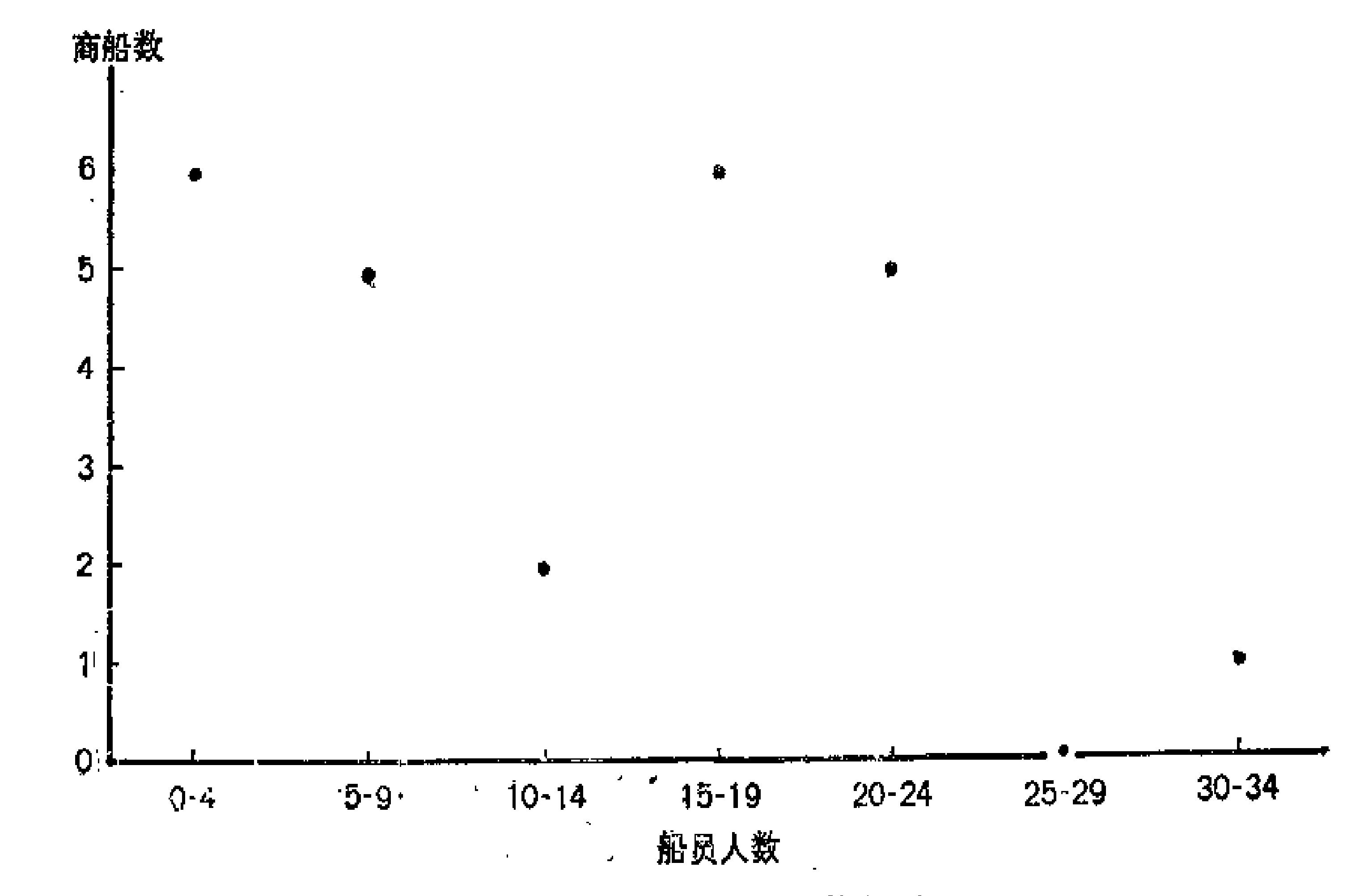


图 4.3 根据表 4.1 第 5 列资料的图表

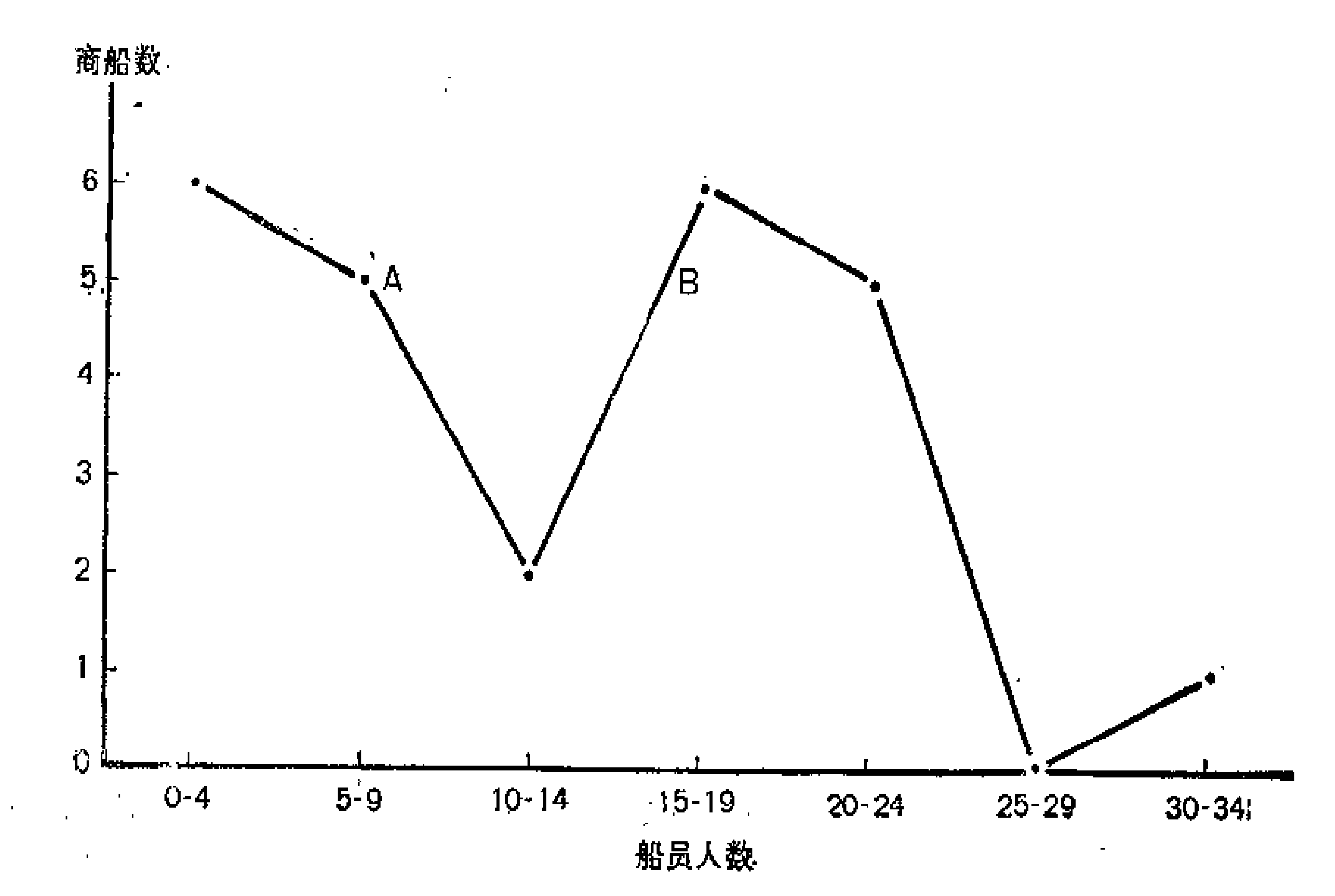


图 4.4　根据表 4.1 第 5 列资料的曲线图

能说有5艘船的人数在10-14和15-19之间的某处，所以B点毫无意义。

曲线图在呈现时间数列资料方面对历史学家特别有价值，因此我们还将回过来讨论它。

当资料是定序或区间类型时，我们也可以应用图解法来呈现交叉分类的结果。我们通过建立一种在统计分析中相当重要的图解，散点图，来这样做。图4.5就是这样一个散点图，其中显示一种对吨位和船员人数所进行的交叉分类。图中每一点代表一艘船，点的位置由船在水平轴上的变量值(吨位)和在垂直轴上的变量值(船员人数)来确定。在本例中，由于只涉及两个变量，哪个应表示在水平轴上和哪个应表示在垂直轴上不成多大问题。

在绘图时有一些基本原则，如果忘记或忽视它们，可能导致解释上的严重错误。图4.6显示，通过加长或缩短图的任何一轴，我们怎样会以导致误解的方式呈现资料，不是突出就是低估波动；为了避免这样，常用的粗略办法就是图的垂直轴应为水平轴长度的三分之二。其它规则是，应始终给出图中的零值，以及应沿着轴把区间明白表示出来。总之，我们应以这样的方式建立图表，即它能够清楚地显示出我们想要突出的资料的特征，而不是采取令人误解的方式。

4.4 比率尺度图

在上一节我们所描述的所有图中，尺度都建立得使水平轴和垂轴上每一类都具有相同的间隔。若我们以表示1770—1800年间英国进口原棉情况的表4.12和图4.7作为一个

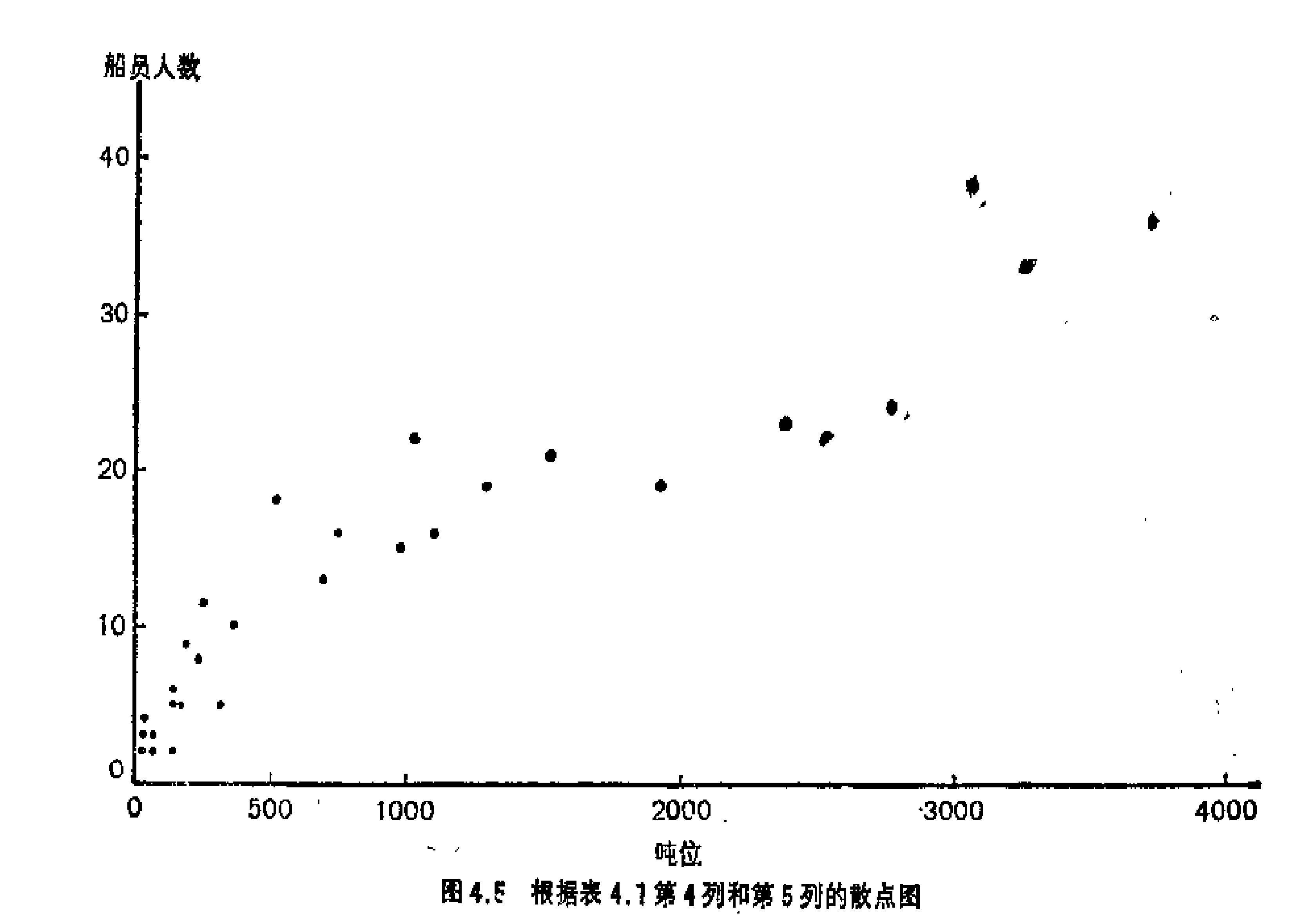


图 4.5　根据表 4.1 第 4 列和第 5 列的散点图

表 4.12　1770—1800年间英国原棉进口的重量

年份	进口量	年份	进口量
1770	3612	1786	19475
1771	2547	1787	23250
1772	5307	1788	20467
1773	2906	1789	32576
1774	5707	1790	31448
1775	6694	1791	28707
1776	6216	1792	34907
1777	7037	1793	19041
1778	6569	1794	24359
1779	5861	1795	26401
1780	6877	1796	32126
1781	5199	1797	23354
1782	11828	1798	31881
1783	9736	1799	43379
1784	11482	1800	56011
1785	18400		

资料来源：B. R. 米切尔和 P. 迪恩：«英国历史统计摘录»(B. R. Mitchell and P. Deane, Abstract of British Historical Statistics), 剑桥：剑桥大学出版社，1962年，第177—178页。

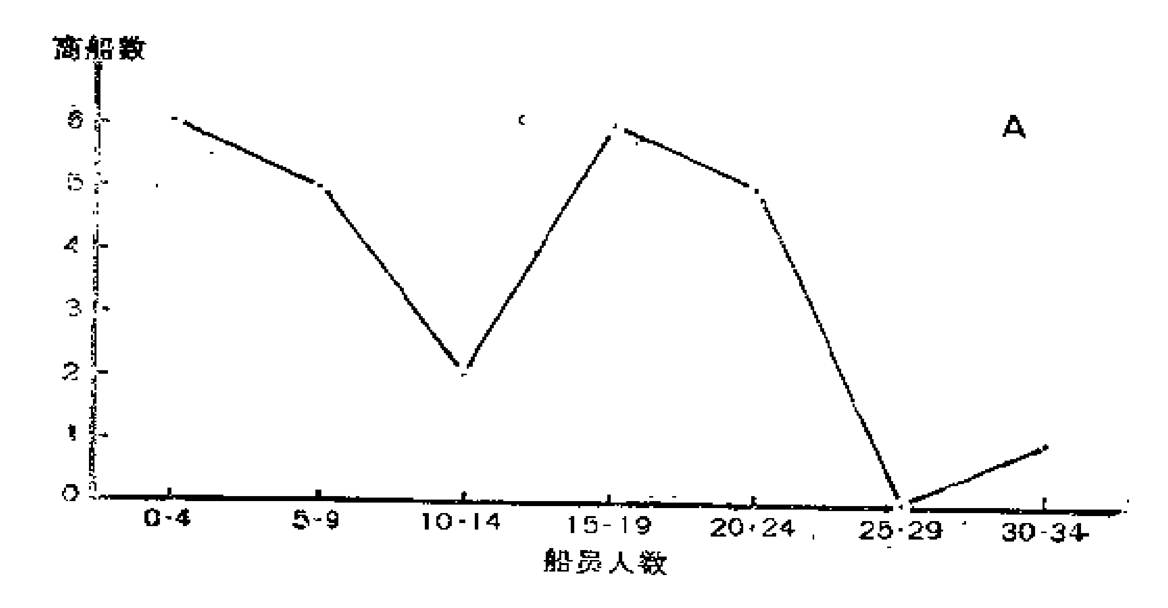


图 4.6-A

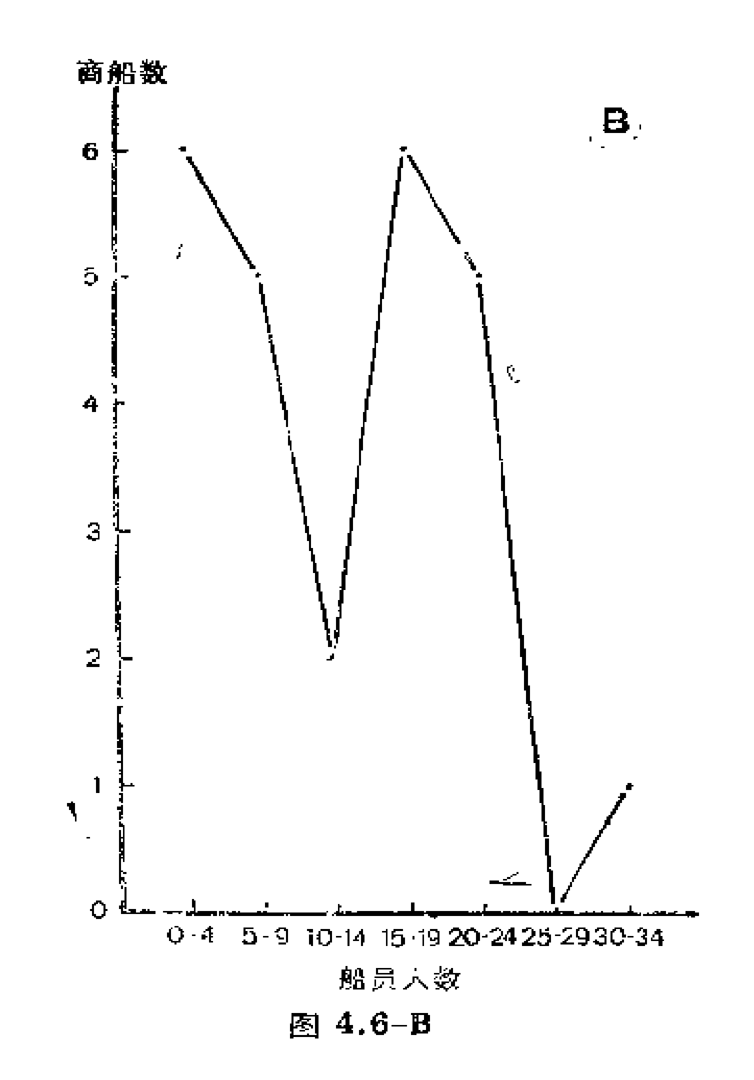


图 4.6-B

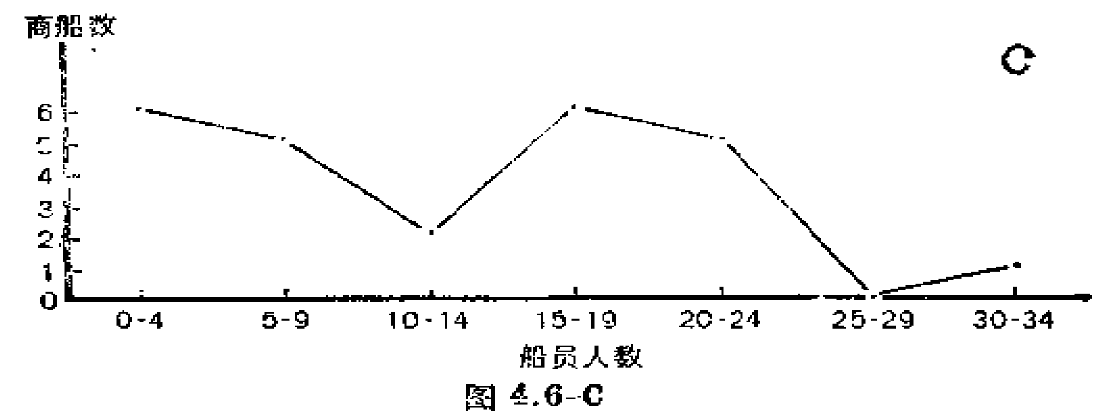


图 4.6-C

图 4.6　改变轴的长度及曲线图区间宽度的影响，根据表 4.7 第 5 列的资料

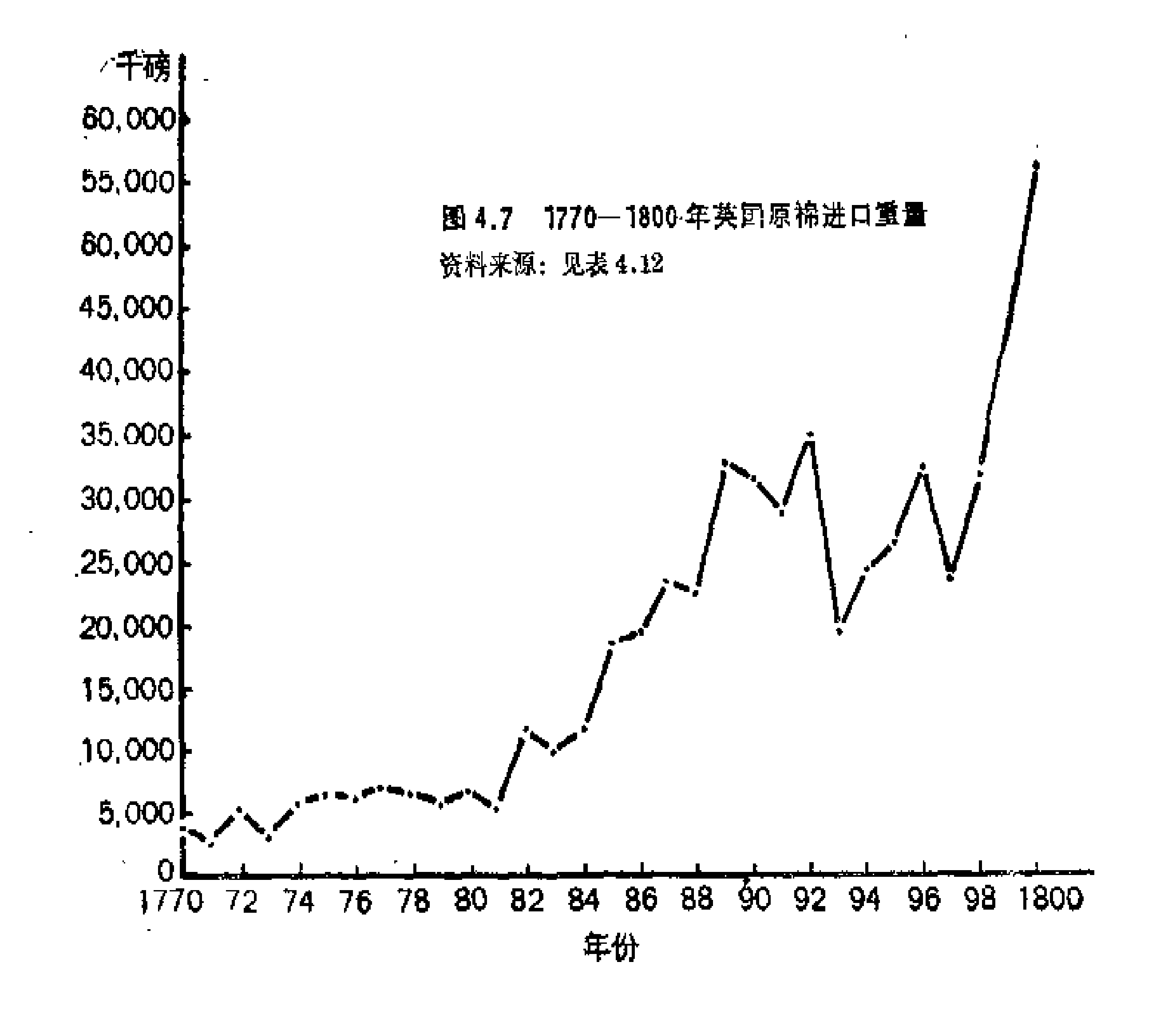


图 4.7　1770—1800 年英国原棉进口重量

资料来源：见表 4.12

进一步的例子，我们见到1772年进口原棉 5,307,000 磅，1773年进口 2,906,000 磅——一年间下跌了 2,401,000 磅。在以后的一个时期，即1790—1791 年间，原棉进口从 31,448,000 磅减至 28,707,000 磅——下跌了 2,631,000 磅。在这两个案例中绝对的下跌值相近似，2,401,000 磅和 2,631,000 磅，因此在图 4.7 的垂直轴尺度上这两个下跌值表示得几乎相等。

然而，在历史研究中，我们常对两个时期之间的相对变化比对其绝对变化更感兴趣。当我们涉及诸如变化很快的英国工业革命时期或涉及像棉花加工对英国那样重要的产业的迅速发展时尤为如此。我们常想要考虑相对增长，比例的或百分比的变化，并比较不同时期的百分比变化；如果我们以1772—1773 年和 1790—1791 年的原棉进口为例，便会发现 1772—1773 年的进口下跌了 45.24%，而 1790—1791 年仅下跌了 8.41%。我们不可能从像图 4.7 中得到这种信息，因为那张图是以同等的绝对变化来表示，而不管百分比变化曾是怎样。因此我们如对用图来表示百分比变化感兴趣，就需要找到另一种形式的图。事实上我们需要一种容易绘制的图，它沿着水平轴和垂直轴给出变量值（这样如果需要就能计算绝对变化），而且按图中相等尺度的区间来表示比例或百分比变化。我们能应用通常基于对数的比率尺度来达到此目的。

从上一章中知道，以10为底的对数具有这样的性质，即10的对数为1.0000，100的对数为2.0000，1000的对数为3.0000。从10至100，亦即增长了10倍，用对数表示就是从1.000增至2.0000，差分为1.0000。同样，从100到1000的10倍增长在对数中也用差分1.0000来表示，尽管从100 到1000 的绝对变化（1000 − 100 = 900）比10到100的绝对变化（100 − 10 = 90）大得

多。

因而对数有这样的性质，即相同的比例变化可用对数中相同的绝对变化来表示，我们希望在图解方法中具有这种性质。因此如果我们将表 4.12 中的每一个值都转换成对数值形式，并将结果标绘在一个垂直轴显示对数的图中，那么我们就已达到了表示比例变化这个重要目的。图 4.8 就是这样的图。

然而，将表 4.12 中的每一个值都变换成对数值仍然是一项很麻烦的操作，况且这样做时我们亦失去了在图中显示原始值的能力。为了从图中便能找到原始值，我们需用反对数表。运用像图 4.9 那样在图中设置的尺度便可以克服这些困难。图 4.9 中尺度上的值都是原始值，而尺度上各点之间的距离表示这些值的对数差分。比如，在尺度上 500 万至 750 万(增长比例为50%)之间的距离与具有相同增长比例的1000万至1500万之间的距离相等。

虽然绘制这类比例尺度很容易，但是购买印有对数尺度的绘图纸却更为方便。然而在买这些绘图纸之前，我们需要预先了解所要表示的值域。对数按周期排列：1.0000—2.0000是第一个周期，2.0000至3.0000是第二个周期，以此类推，而绘图纸是为表示 1 个，2 个，3 个或更多的周期而设的。假如我们想要在图中包括 3—1750 的资料，就应购买设有4个周期的绘图纸：1-10(log 0.0000—1.0000)，10-100(log 1.0000—2.0000)，100-1000(log 2.0000—3.0000)，1000-10000(log 3.0000—4.0000)。

还应指出，图 4.8 恰当地被称为半对数图，因为只有 1 个轴上标有对数尺度；用对数尺度标绘时间变化通常没有什么

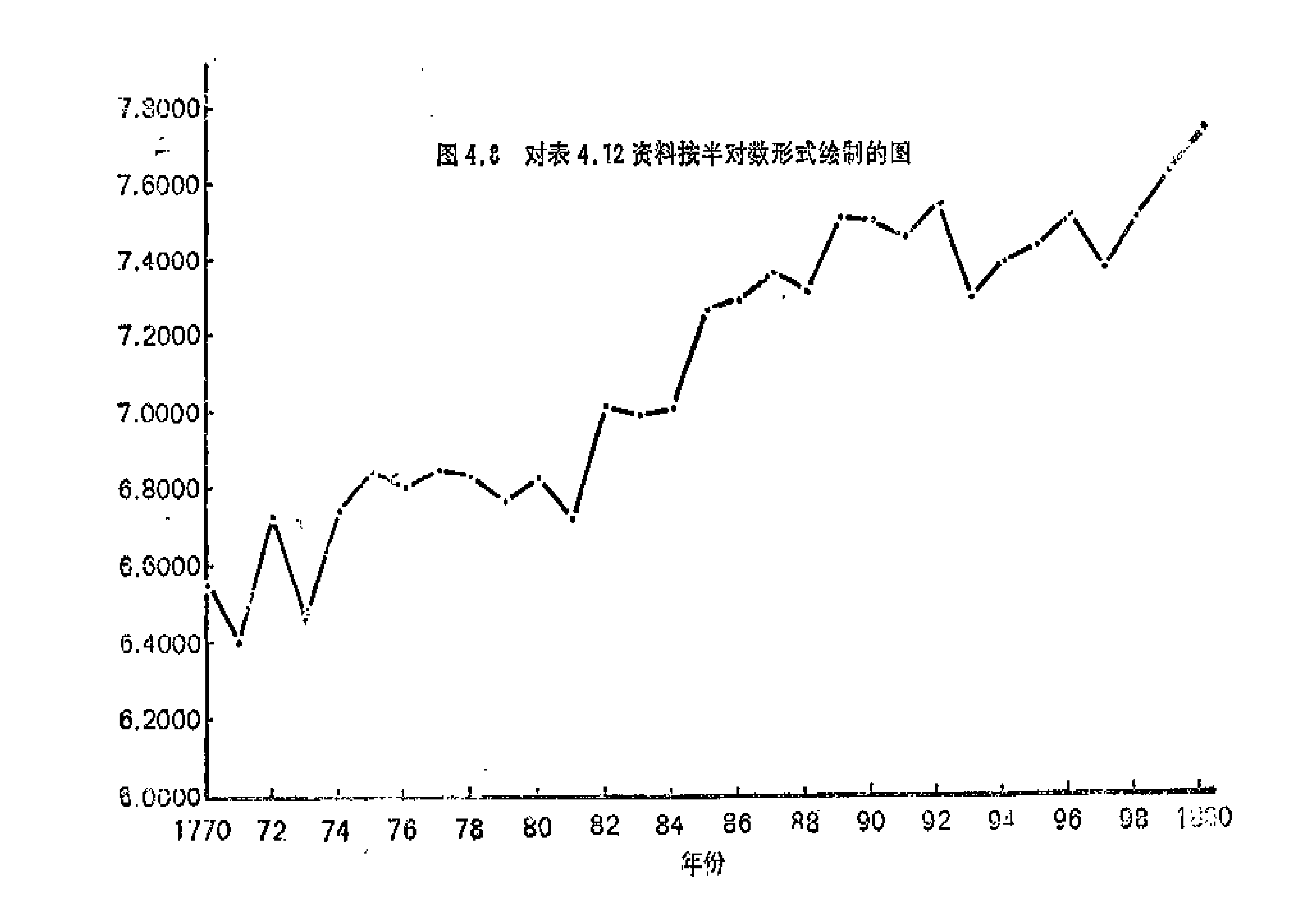


图 4.8　对表 4.12 资料按半对数形式绘制的图

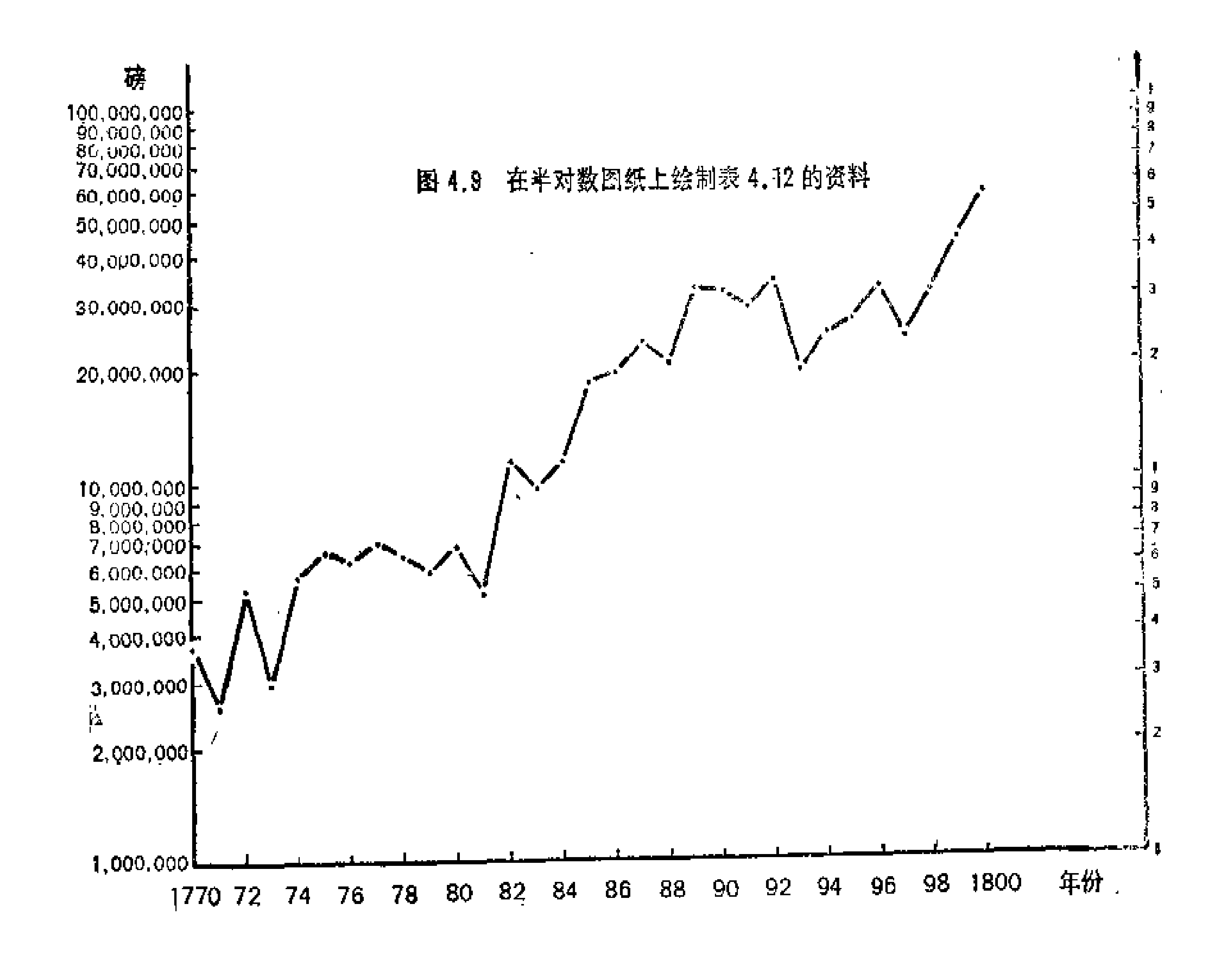


图 4.9 在半对数图纸上绘制表 4.12 的资料

意义。若我们打算为2个对应的变量绘图，并考察每一变量的比例变化时，可以用对数图，在其中两种尺度都用对数标记；这种绘图纸也能买到。

对半对数图和对数尺度图的解释需要谨慎，因为我们的眼睛习惯于具有相等区间尺度的图。应记住，我们要特别注意图中两点间连线的斜度：斜度越陡比例变化越快。在第六章讨论时间数列资料时我们将再来谈谈对数图和相关的对数变换的进一步应用。

5 资料的初步分析II：概括性方法

上一章讨论了可以用来对一个原始资料矩阵进行重新整理和呈现，以增加我们对资料所含信息的理解的一些方法。虽然在建立编组频数分布时也曾试图对资料加以概括，但对此并未进行得很深入；我们仅把资料归并为较少的几类。随着频数分布的编组数的减少，资料亦丧失其一定界限，因此这种方法也不可能推行得过于深入。如果我们只想得到1个能充分概括资料的数字，那就不可能使用频数分布；因为只有1个编组的频数分布仅能告诉我们资料矩阵中的个案数。

因此在这一章里，将讨论对我们已遇到过的不同类型的资料进行概括的其他方法。应该指出，我们将把迄今在说明中所用的次序颠倒过来，从适用于区间资料的方法开始，然后再转向适合于定序和定名资料的概括性方法。

5.1 算术平均数

算术平均数方法就是一种将使我们能计算出1个数目来表示或概括一整集数目的方法。算术平均数更以“平均数”

著称，但这种说法易引起误解；下面还要谈到其他种类的平均数，所以还是为求准确并应用"算术平均数"这个名称为好。这种方法仅适用于区间资料。

将一纵列中的数目累加起来再除以个案的数目便很容易算出算术平均数。例如，根据表 3.1 中的资料，表里给出的牧猪总数是11,915；有 50 个个案，因此算术平均数为11,915/50＝238.3头牧猪。我们很容易运用求和记法并以符号$\overline{X}$（读作X横）来指示向量X的平均数，其公式为

$$\overline{X}=\frac{\sum_{i=1}^{N}X_i}{N}$$

（可能有人异议，在表 3.1 中的资料中，似乎牧猪数目总是四舍五入到最接近于 5。在这类案例中，平均数的计算可能有偏差，但是四舍五入到最接近于 5 而不是更大的区间，误差很可能不大，计算平均数的程序完全可以接受）。

从已经重新整理成频数分布的资料而不是原始资料中计算某些资料的平均数常是合适的。这种计算只是稍微复杂一些，然而它所花费的额外工作因个案数目较少而得到补偿。对一个频数分布计算平均数，需用频数分布的每一个值乘以它所对应的频数，将其结果求和，再将总数除以原始资料中的个案数目。例如表 5.1 即是计算表 3.2 资料的平均数。其结果238.3恰恰与用原始数据得到的结果相一致。在一般情况下，一个频数分布的平均数由下面的等式得出

$$\overline{X}=\frac{\sum_{i=1}^{k}f_iX_i}{N}$$

表5.1 根据频数分布计算平均数

牧猪头数 X_i	地区数 f_i	f_iX_i
5	1	5
15	1	15
20	2	40
30	9	270
40	1	40
50	2	100
55	1	55
60	4	240
80	2	160
100	2	200
120	1	120
150	3	450
160	2	320
200	3	600
300	1	300
350	2	700
400	3	1200
500	3	1500
600	3	1800
800	2	1600
1000	1	1000
1200	1	1200
	N＝50	11915

$$\frac{\sum_{i=1}^{k} f_i X_i}{N}=\frac{11915}{50}=238.5$$

（这里 k 是编组数）

这里，X_i是变量的值，f_i是这些值出现的频数，k为编组数，而N则是编制频数分布所依据的个案数目。

在表3.2中，有关每一教区牧猪数目的资料被呈现为一个频数分布，但未对资料进行概括。然而在表3.3和表3.4中，资料被归并成组或类，这种类型的频数分布因而被称为编组频数分布。当有必要根据这类编组频数分布来计算平均数或其他统计量时，需要一不同的程序，这种程序考虑到给出的不再是资料的真值，而仅是它们所属编组这一事实。

为从编组资料中计算平均数，我们假设每一个案处于它所属编组的中点。为了计算出这个中点，我们必须仔细地检查曾经用过的编组方法。例如，在表3.4中我们把编组规定为0—99头牧猪，100—199头牧猪，等等。这样做时，我们并没有考虑到当碰到一个具有99.7头牧猪的教区时该怎么办；由于0.7头牧猪是一个没有意义的数量，我们不需考虑这种可能性。但是，假如我们使用其他资料的话（如《末日裁判书》中有关教区土地面积的资料），可能会发现一个拥有99.7英亩土地面积的教区，把这个教区分配到任何一个编组都会成为问题。通常我们将数字四舍五入到最接近的整数，以此将面积为99，99.1，99.2，99.3和99.4英亩的土地都归于第一编组（0—99英亩），而将99.5，99.6，99.7，99.8和99.9英亩的土地则归于100—199英亩那一编组。因此，任何低于99.5英亩的土地都被归属于较低的编组，而任何高于99.5英亩（包括恰好为99.5英亩）的土地则被归属于较高的编组。而而虽则在表3.4中我们规定各编组的范围是从0—99和100—199，而实际在-0.05—99.5，99.5—199.5等值之间；由于这个原因，这些组的极限值被称为"真极限"（与"规定极限"相对），而

在计算编组频数分布中编组的中点时采用的就是这些"真极限"。回到表 3.4 中的资料，我们便可以再现具有真极限、说明极限和编组中点的表格。表 5.2 以实例说明了根据这类编组频数分布计算平均数的方法。

表5.2　根据编组频数分布计算算术平均数

牧猪数目规定极限	牧猪数目真极限	编组中点 m_i	个案数 f_i	$m_i f_i$
0-99	—0.5-99.5	50.0	23	1150
100-199	99.5-199.5	150.0	8	1200
200-299	199.5-299.5	250.0	3	750
300-399	299.5-399.5	350.0	3	1050
400-499	399.5-499.5	450.0	3	1350
500-599	499.5-599.5	550.0	3	1650
600-699	599.5-699.5	650.0	3	1950
700-799	699.5-799.5	750.0	0	0
800-899	799.5-899.5	850.0	2	1700
900-999	899.5-999.5	950.0	0	0
1000-1099	999.5-1099.5	1050.0	1	1050
1100-1199	1099.5-1199.5	1150.0	0	0
1200-1299	1199.5-1299.5	1250.0	1	1250
			$N=50$	13100

$$\overline{X}=\frac{\sum_{i=1}^{k} m_i f_i}{N}=\frac{13100}{50}=262.0$$

如表 5.2 所示，根据编组频数分布计算算术平均数得出的结果为 262.0。而根据原始资料的全部向量得到的结果为 238.3。正确的算术平均数 238.3 与根据编组资料计算的平均数262.0之间的差，即是我们为应用较为便利的编组频数分布在最终结果的不精确性上所付出的代价。根据原始资料得到的平均数与根据编组资料得到的平均数相异的差距，取决于

实际资料与所选编组的中点相异的程度和范围而定。这强调了选择恰如其分的编组的重要性，特别对编组资料的计算尤为如此。

与我们将要讨论的其他概括方法相比，平均数容易计算并有这个优点，它不仅考虑到一个分布中项目的数量，还考虑到每一项目的值。算值平均数的一个相应的缺点是，正是由于它包括每一个值，一个极端值的存在会对它产生相当大的影响。

让我们以表 4.1 第 4 列给出的有关 25 艘商船吨位的资料为例。商船的总吨位 $\sum X_i$ 在表 4.8 中算出是 22,319 吨。除以 $N=25$，这得到一个 892.76 吨的吨位算术平均数。然而，其中一艘船的吨位比任何其他船都多 500 吨以上（官方登记号为99437）。如果我们从平均数的计算中剔除这艘船，总吨位数将减至19,073吨，吨位平均数将降为794.71吨。因此，把这艘船包括进去很大地影响了平均数。同样，若包括一艘仅有 26 吨的小船（登记号为95757）也将会使平均数降低许多。

部分地是由于这个原因，我们需将平均数与其他一些平均数从中算出的资料的值域的测度联系起来，因此我们下面就讨论这样一种测度，标准差。

5.2 标准差

无疑，平均数是概括区间资料最简单和最合适的方法。它设计得将人们的注意力集中于被考虑的一组资料中的所谓“集中趋势”上。在仅有 2 个数目需加计算的最简单例子中，我们可以想象有一条线，上面有 2 个点代表这 2 个数；于是代

表平均数的第三点如图 5.1 所示将处于这 2 点之间的中心位置。原始数字用 A 和 A' 表示，平均数用 $\bar{X}$ 表示。

如果我们在这条线的两端离 A 和 A' 距离相等的相反方向上再置 2 个点，B和 B'，并考虑结果会怎样，运用平均数作为一种概括性方法的主要困难就显示出来了。现在，让我们找出 B 和 B' 所代表数目的平均数；由于从 A 到 B 和从 A' 到 B' 的所距相等，很明显 B 和 B' 的平均数也将是 $\bar{X}$。相类似地，如果我们从 A 和 A' 向中心以相等的距离再置 2 个点，C 和 C'，它们的平均数还将是 $\bar{X}$。

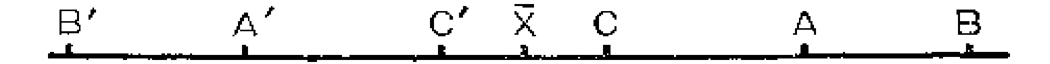


图 5.1 算术平均数的图形表示

因此，平均数并未指出资料的逐个观察值与平均数的偏离程度。所以如果我们打算应用算术平均数来概括资料，还要运用其他一些方法来描述围绕平均数资料的离中或相异的量。做这个的最简单的方法似乎是将每一观察值与平均数相异的量合计起来，这一量被称为每一形成分布一部分的观察值与分布的平均数的“偏差”。

然而，这并不是解决表示围绕平均数的离中趋势问题的正确方法，因为这种计算的结果总是零。从对图 5.1 的观察中就可以看出，这是由于对平均数的单个偏差将会相互抵销。因此，有必要发展另外一种能避免这一缺点的离中趋势的测度方法。

正如集中趋势的测度除了算术平均数外还有其他方法那样，离中趋势的测度方法也有多种。我们将把其他方法放到本书的以后部分讨论，而现在只集中讨论最为方便并且最为广

泛运用的离中趋势测度方法，它就被称为"标准差"。与平均数一样，只有当资料是区间或比率类型时才能计算标准差。

我们刚讨论过的表示离中趋势方法的困难在于，所有大于平均数的观察值都被小于平均数的观察值所抵销。不顾偏差的符号，计算其总数，再除以观察值的个数就得到偏差绝对值(没有符号的值)的平均数，这样就可以避免上述困难。其结果被称为平均偏差，它有时被用于统计，但它也有缺点；特别是计算起来很费事，当观察值的数目很大时尤为如此，并且不能用于进一步的分析。我们另外可以通过把每一偏差平方而去掉偏差的负号来避免上述困难。大家记得，无论是一个正数还是一个负数的平方都将是一个正数。由于我们感兴趣的是围绕平均数的平均离中趋势，我们在计算所有平方偏差之和，然后除以项目的数目。对于向量X，其公式为

$$\frac{\sum_{i=1}^{N}(X_i-\bar{X})^2}{N}$$

其结果被称为"方差"。

表5.3中，我们取表3.1牧猪头数那一纵列的前10个个案为例来说明计算方差的过程。方差虽然容易计算并在以后的工作中很有用处；但作为一种离中趋势测度方法，它有两个缺点。其次要的缺点是：如果从平均数得出的偏差值大，对它们进行平方就会使它们变得更大，处理起来就更麻烦。主要的缺点在于很难对方差给出一个实际的，而不是数学上的意义。还以表5.3中的资料为例，说1086年埃塞克斯10个地方牧猪头数的算术平均数为348头十分直截了当；若说这一平均数的平均离中趋势是104,896.0平方头牧猪就没有任何意义。

部分地是由于这个原因，部分是由于围绕平均数的离中趋势可以用与平均数本身相同的单位来表示，我们才应用最便利的离中趋势测度方法，即标准差——就是方差的平方根。对于向量 X，它是

$$\sqrt{\left(\frac{\sum_{i=1}^{N}(X_i-\overline{X})^2}{N}\right)}$$

通常用字母 S 表示。在表 5.3 的例子中

$$S=\sqrt{104896.0}=323.9$$

表 5.3　方差的计算

地　　区	牧猪头数 X_i	平均数 $\overline{X}$	$(X_i-\overline{X})$	$(X_i-\overline{X})^2$
里特尔(Writtle)	1200	348	+852	725904
克拉弗林(Clavering)	600	348	+252	63504
法恩哈姆(Farnham)	150	348	−198	39204
法恩哈姆(Farnham)	50	348	−298	88804
乌格莱(Ugley)	160	348	−188	35344
阿尔费勒斯图那(Alferestuna)	350	348	+2	4
肯菲尔特(Canfield)	120	348	−228	51984
邓莫(Dunmow)	300	348	−48	2304
伊斯顿(Easton)	150	348	−198	39204
伊斯顿(Easton)	400	348	+52	2704

$N=10$　　$\sum_{i=1}^{N}X_i=3480$　　$\sum_{i=1}^{N}(X_i-\overline{X})=0$

$$\sum_{i=1}^{N}(X_i-\overline{X})^2=1048960$$

$$\frac{\sum_{i=1}^{N}(X_i-\overline{X})^2}{N}=\frac{1048960}{10}=104896.0$$

而因此我们可以说表 5.3 中资料的标准差为 323.9 头牧猪。

从表 5.3 中可以看出，为了计算标准差首先须计算平均

数。我们常需计算平均数，然后再能用上面的公式计算标准差；但是我们还可以从其他公式直接计算标准差，这些公式都是根据标准公式重新整理出来的。其中最实用的公式是

$$S=\frac{1}{N}\sqrt{\left(N\sum_{i=1}^{N}X_i^2-(\sum_{i=1}^{N}X_i)^2\right)}$$

它仅包括变量值总和及每一值平方之和的计算。其他的公式以及从未编组的和已编组的频数分布中计算平均数的公式，在任何统计教科书中都可找到。

在我们处理区间资料时，计算平均数和标准差有很多优点。事实上，算术平均数是本书将要讨论的各种“平均数”中最为人所熟知和用得最多的。算术平均数和标准差的主要优点在于计算中资料的每一项都被应用。因此没有浪费任何信息；况且，两种方法无论是通过手工还是通过计算机器都容易计算。正如我们曾经谈到的，算术平均数的缺点在于它对频数分布中极端值的存在很敏感，这一点在历史问题中很重要，当涉及到财富和社会地位悬殊很大的前工业社会时尤为显著。例如，若我们对计算一个中世纪村庄中家庭人口的平均数感兴趣，或许会发现由于村里存在着一个城堡，其中住有贵族及他的仆人和家臣，这个平均数趋于偏高；在这个例子中，平均数使人对正常家庭人口数目产生错误的印象。

这一困难，通常并不像上述例子中那样明显，似乎可以通过把标准差与平均数相联系，作为资料离中趋势的测度方法而被克服。不幸的是，在一个频数分布中极端值——也就是与大多数其他值相距甚远的值——的存在也影响到标准差的值。大家记得，标准差的计算中包括将每一个平均数偏差值平方。尽管我们试图运用开方对此进行补偿，与接近平均数的

值相比较，一个远离平均数的值不可避免地要被给予较大的分量。当我们讨论较高级的统计方法时就会知道，这一点并非总是缺点，但是当算术平均数和标准差被考虑作为概括性方法时，我们要时常记住这个困难。在本章的以后部分还要讨论集中趋势和离中趋势的其他测度方法；应用其中的某些测度方法有助于解决这一困难。

5.3 几何平均数

几何平均数是仅适合于区间资料的第二种平均被，除了在经济学和经济史的问题中，它很少被应用。对于一个总数为N的向量 X，计算几何平均数的公式是

$$G.M.=\sqrt[N]{[(X_1)(X_2)(X_3)\cdots(X_N)]}$$

换句话说，我们将所有N个X值相乘，再开N次方。在实践中，我们可以应用对数，将所有N个X值的对数相加，其结果除以N，再取反对数。在某些情况下，尤其是处理指数及要找到在一个变量或许多变量中的相对变化的平均数，如想找到若干增长率的平均数时，几何平均数是最合适的测度集中趋势的方法。与算术平均数相比较，它往往给概端值以较小的分量，但没有任何离中趋势的测度方法直接与它相联系。

5.4 中位数

第三种集中趋势的概括性方法为中位数，它可以用于定序或区间资料的计算。像下面的众数一样，中位数的计算十

分简单；它只是将一个有序排列的个案分为两半，并使小于中位数值的个案数与大于中位数值的个案数相等的一个变量值。因此，为计算中位数我们仅需按照个案对某一特定变量取得的值将它们顺序排列；中位数就将是位于这个次序半中央的值。例如，我们取表 4.1 中给出船员人数的第 4 列，在表5.4中重新将个案排序。

由于表 5.4 中个案的数目为奇数(25)，处于中位的个案数为第 13$(N+1)/2$；因此这个分布的中位数值正巧也是 13。如果分布的个案数为偶数，比如说 24，那么中位数则取第12个个案 $N/2$ 与第13个个案$(N+1)/2$的算术平均数；以表5.4为例(假设第 25 个个案已被排除)，其余的 24 个个案的中位

表 5.4　按照顺序对表 4.1 中第5列资料的整理

序号	1	2	3	4	5
官方号码	113685	114424	115143	1697	113689
船员人数	2	2	2	3	3
	6	7	8	9	10
	95757	35052	93086	2640	62595
	4	5	5	6	8
	11	12	13	14	15
	123375	107004	94546	86658	73742
	9	10	13	15	16
	16	17	18	19	20
	109597	115149	96414	99495	115357
	16	18	19	19	21
	21	22	23	24	25
	113406	114433	92929	118852	99437
	22	22	23	24	33

数将是第12与第13个个案的平均数：

$$\frac{10+13}{2}=11.5$$

中位数仅是按照值的排列划分和概括资料的一组测度方法之一。除了中位数，还有四分位数，十分位数和百分位数。正像中位数将个案分成2组，四分位数将资料分成4组，十分位数为10组，百分位数为100组。不幸的是，有两个计算四分位数的惯例；最常见的一个即是第一个四分位数被规定为第〔(N+1)〕/4位个案，第二个四分位数(中位数)被规定为第〔(N+1)〕/2位个案。第三个四分位数则是第〔($3N$+3)〕/4位个案(“较低的”和“较高的”有时被用来代替“第一个”和“第三个”)。按照这个规则，表5.4中例子将得出6.5，13，19.5，经四舍五入后我们得到作为四分位数的第13，第17和第20位个案，它们的值分别是5，13和21。十分位数和百分位数的计算与此相似，当然它们的应用限于资料集是由一个数目很大的个案所组成，致使把资料分成这样多的组是有意义的情况。

各种离中趋势测度方法可以与四分位数、十分位数和百分位数联系起来应用；最常用的是四分位数偏差，或者更确当地是半内四分位数距，即是第1与第3个四分位数差的一半。因此，在表5.4的例子中，半内四分位数距是

$$\frac{21-5}{2}=8$$

因此，对这个例子我们可以说中位数为13、半内四分位数距为8，使我们对分布的集中趋势和围绕这个集中趋势的离中趋势的量有了一些概念。

虽然中位数和四分位偏差很容易计算并且是很方便的集

中趋势和离中趋势的测度方法，它们也有一些缺点，使我们在通常的情况下有其他选择时就不用它们，如我们处理区间资料时就是那样；若资料属定序类型时，我们仅有两种选择：众数和中位数。中位数和四分位偏差最大的缺点在于，它们在计算中没有考虑到分布中的极端值，而只是表明它们存在。例如，让我们假设用表 5.5 中的资料代替表 5.4 中已经呈现过的真实资料。这个分布从形状上看与表 5.4 有明显的不同，而且在表 5.4 和表 5.5 的等级次序中仅有 6 项具有相同的值。再则表5.5中等级次序最高的极端值比表5.4的最高极端值要大得多。可是两项分布却有着相同的中位数，相同的四分位数偏差，而且即便我们改变假设的分布使第 20 位以上的所有值翻几倍依然如此。应该指出，在某些情况下不顾那些与频数分布中大多数值相差甚远的值的集中趋势测度方法也有优点；例如在我们研究人们结婚的正常年龄时，不顾 55 岁才结婚的老处女（比大多数同代人晚婚 30—35 年）的一种测度十分有用。与此类似，如我们对一工业城镇中工人生活水准感兴趣，实际工资中位数将比实际工资平均数对生活的物质水准提供更好的概念；中位数几乎不受生活在这一地区的工厂主的实

表5.5　一组25个假设数目的排列次序

序号	1	2	3	4	5	6	7	8	9	10
值(船员人数)	2	2	2	2	2	2	2	4	6	7
	11	12	13	14	15	16	17	18	19	20
	7	9	10	10	10	11	12	13	16	20
	21	22	23	24	25					
	38	69	77	95	160					

际高收入的影响。然而，对大多数资料来说，中位数对极端值的不敏感是一个缺点。需要强调指出，中位数的再一个缺点是，仅有很少的统计分析方法应用它。一般说来，如果应用的是区间或比率型资料，中位数并不是一个很有用的集中趋势的测度方法，而与之相联系的离中趋势测度方法，即四分位数偏差，也仅用于少数特殊事例。

5.5 众数

如果我们有表 4.1 所示资料矩阵中第2和第3列那样的定名资料，那么唯一可以采用的集中趋势的概括性方法就是众数。众数仅是最经常出现的那个值。在给出商船的动力方式的变量的例子中，从表 4.2 我们知道有 4 艘船用风帆、16艘用蒸汽驱动，5艘没给出推进方式。因此，变量"推进方式"的众数值为"蒸汽"。众数也可以被用作为一种定序或比率型资料的概括性方法；表 4.1 的变量"船员人数"的众数是 2，因为有3艘船拥有这个数目的船员，而其它的船员人数出现不超过2次。

很明显，从这些例子中可以看出众数仅限于概括表 4.1 中那样的资料；这一点在众数应用于历史资料时的确如此，虽则在一些事例中了解一个资料集中最共通的值十分重要。例如，由于大多数人结婚或生第一个孩子的年龄上的变化会影响到出生率，人口史学家常应用众数。但是众数的主要缺点在于它没有与之相联系的离中趋势测度方法，因此当资料稍属离散型时众数的应用受到很大限制。

5.6 变异系数

在对历史学家很可能有用并为他们所用的概括性测度方法中只有变异系数了。当资料是区间类型时，变异系数提供了一个比较两组变量的各自平均数差异程度的简单手段。它对了解离平均数最远的2个或3个变量用处很大；比如，在商船的例子中，我们或许有兴趣知道商船吨位间的不同是否甚于船员人数间的不同。部分是由于它们的计量单位不同（人数和吨数），部分是由于它们的平均数相差甚大，我们不能直接比较每一组变量的标准差。任何向量数的变异系数仅是将这一向量的标准差表示成所占这一向量平均数的百分比。这样，表 4.1 中资料的变异系数为：

吨位 $\overline{X} = 892.8$ 吨

标准差 $S = 946.2$ 吨

$$变异系数 = \frac{946.2}{892.8} \times 100 = 105.99$$

船员人数 $\overline{X} = 12.8$

$S = 8.6$

$$变异系数 = \frac{8.6}{12.8} \times 100 = 67.19$$

因此，我们可以说围绕着各自的平均数，商船吨位较之船员人数显示出更大的变异。

5.7 运用哪一种?

对一个特定的资料集，选择哪种概括性方法首先取决于资料的类型，其次取决于资料的特征，尤其是资料的变异范围，而第三取决于概括性方法在以后分析阶段的应用情况。对于某些案例，这种选择并没有清楚的界限；例如，在人口统计工作中，每一种概括性方法都仅适合于阐述资料的某一特征——比如，众数将给出最共通的结婚年龄，中位数和算术平均数将分别给出“正常”的结婚年龄，中位数在某种程度上排除，而算术平均数则包括大多数非正常的个案。在这些情况中，不同的集中趋势测度方法阐明资料的不同方面，而且它们都能有用地给出不同的方面。基于研究在德文的一个乡村里婚姻情况的表5.6，已做到这点。表5.6显示出应用集中趋势的

表 5.6　考利登乡初婚年龄

男　人	数目	平均数	中位数	众数
1560-1646	258	27.2	25.8	23.0
1647-1719	109	27.7	26.4	23.8
1720-1769	90	25.7	25.1	23.9
1770-1837	219	26.5	25.8	24.4
妇　女				
1560-1646	371	27.0	25.9	23.7
1647-1719	136	29.6	27.5	23.3
1720-1769	104	26.8	25.7	23.5
1770-1837	275	25.1	24.0	21.8

资料来源：E. A. 里格利：‘前工业化时期英国的家庭人员限制’（E. A. Wrigley, ‘Family limitation in pre-industrial England’），«经济史评论»第19期(1966年4月)第1卷，第86页。众数是从平均数和中位数内推，而不是从资料中直接得出的。

一种测度方法而没有其他测度方法，可能会给出错误的印象。例如，单独应用平均数将给人以这样的印象，即在几百年间结婚年龄的变化很大；而另一方面，众数告知我们在几百年里最共通的结婚年龄没有多大变化。只引用两种概括性测度方法之一而不及其他，都会引起误解；因此，里格利引用了所有3种测度方法为我们提供了下结论所需的材料。

了解和运用这3种测度方法之间的相互关系十分重要。在一个简单案例中，平均数、中位数和众数之间的相互关系是清楚的。如果考利登乡中9项婚姻以图5.2所示的方式发生；那么初婚年龄平均数、初婚年龄中位数和初婚年龄众数是一致的；换句话说，在图5.2这类资料的对称分布中，3种测度方法得出相同的值。在图5.2中，它们都是25，我们说这一分布的“高峰”在25岁。我们可以同样设想每项婚姻都发生在两年以前。倘若如此，整个分布应沿水平轴向左移两年，分布的峰以及平均数、中位数和众数将成为23，它的对称性仍将得以保持。

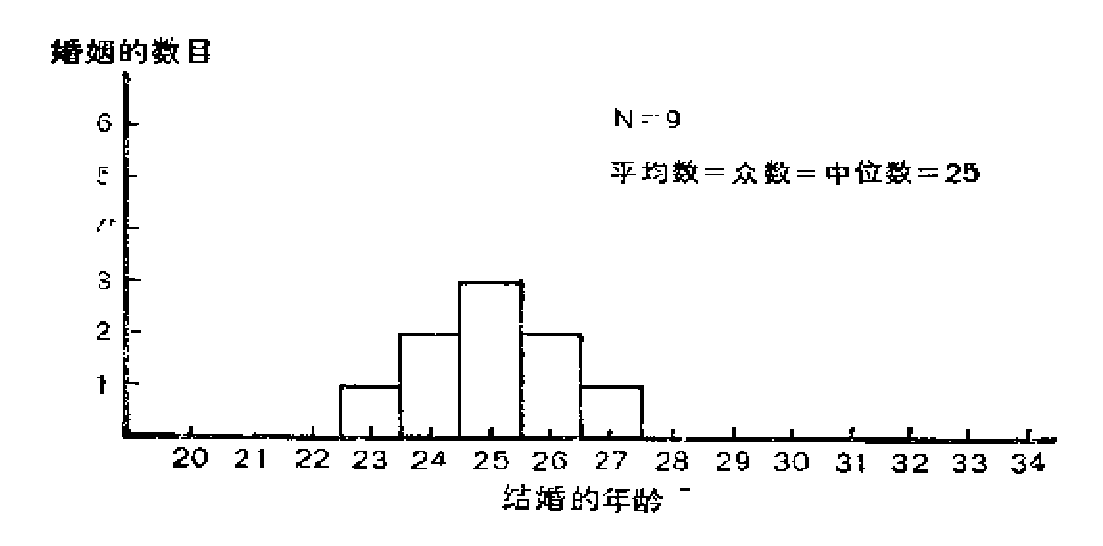


图5.2　对称分布

如果当资料的分布呈对称时，平均数、中位数和众数恰好重合，那么完全合理并正确地推论，如这三者经过计算而发现不一致，如表5.6，那是由于分布不对称所致。例如，在图5.3中11项较大的结婚年龄已加到图5.2所示之中。很明显图5.3中的分布是非对称的，而集中趋势的测度有不同的值。众数仍保持为25，但中位数已升到26；平均数甚至已升到 26.25 。简单的实验表明，如果加入更多的超过众数的结婚年龄，尤其是比众数高得多的年龄的婚姻，平均数和中位数将分歧得更甚，如图5.4。换句话说，我们可以称图5.3这类非对称分布为偏斜分布，并且随着加入更多项的婚姻它会变得更为偏斜。图5.3中，由于超过众数（向右）的婚姻多于低于众数的婚姻（向左），我们说分布向右偏斜；如果其他婚姻的出现低于众数年龄，其分布看起来就像图5.5，并被说成向左偏斜。

具备了上述知识，我们就可以回过头来看表5.6所展示的测度。如果我们看妇女婚姻，概括性测度必然不是产生于对称分布，是产生于偏斜分布。而且，表现1647—1719年的婚姻分布比1560—1646年的分布更向右偏斜，而1720—1769年的分布与早期的分布又极为相似。这三个时期初婚年龄众数几乎保持不变，因此我们可以推断出，考利登乡初婚年龄平均数的变化主要是由于1647—1719年的完全超过众数（或最共通的）年龄的婚姻成比例地增加既多于1560—1646年，又多于1720—1769年所造成的。（我们必须说"成比例地"增加，因为婚姻的数目变化很大。）所以，由于整个分布向右移动了2.6年（29.6—27.0），也由于它变得更向右偏斜，考利登乡的初婚年龄平均数并未改变。

这有两种含义。首先，它改变了我们对作为考利登乡婚

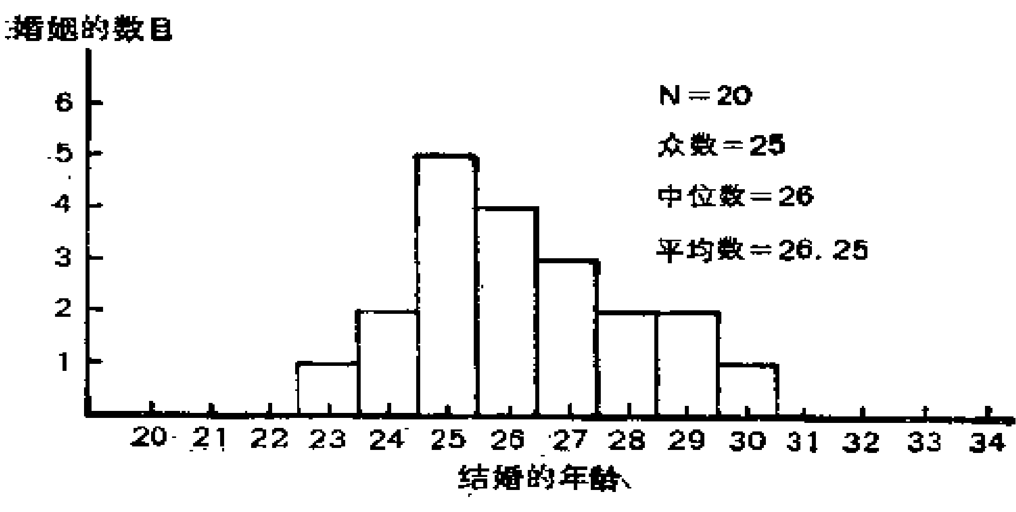


图 5.3 一个向右偏斜的分布

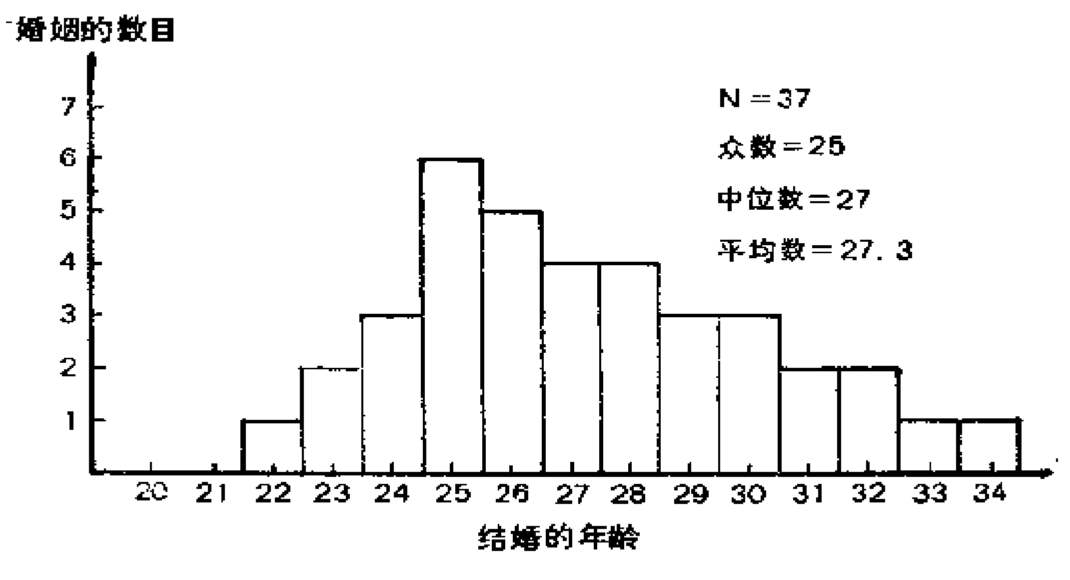


图 5.4 一个向右偏斜更远的分布

姻变化的基础的历史过程的观点和解释。其二，从应用统计学的观点看，它强调测度集中趋势的目的在于总结和阐明从频数分布中得出不同特征这一事实。它们不能代替而只能补充对频数分布的形状以及塑造这一形状的历史过程性质的细心检查。

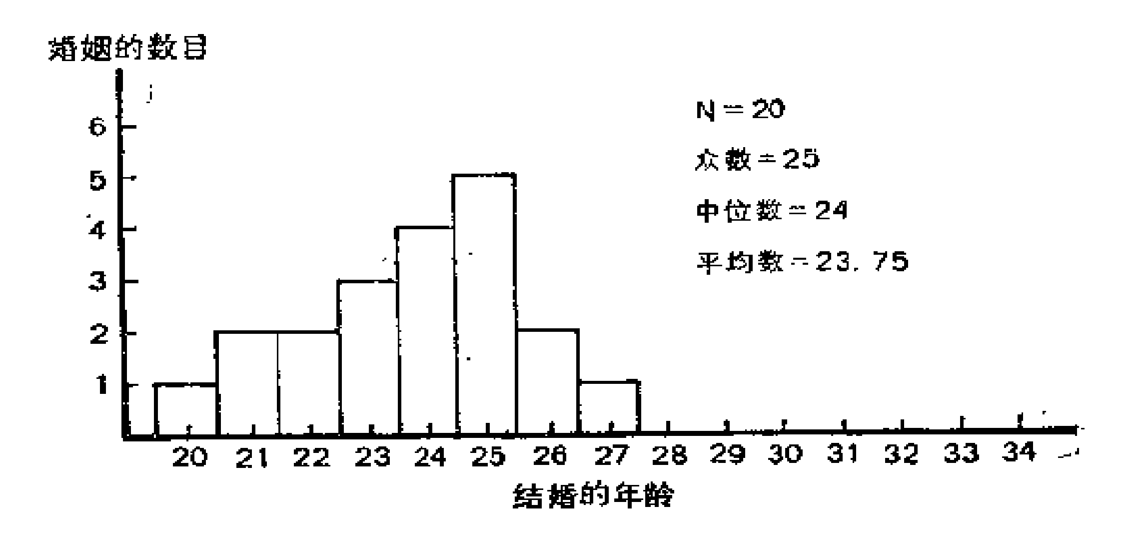


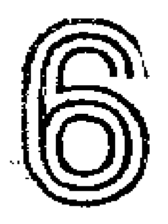
图 5.5　一个向左偏斜的分布

最后一点：表5.6中的众数是用下面的公式从平均数和中位数计算出来的

平均数－众数＝3(平均数－中位数)

这一公式仅适用于轻度偏斜的分布，但如果众数计算起来很麻烦，或者分布呈不规则形状，并像有时发生的那样，有两个以上的高峰，应用它是方便的。

应用所有概括性测度方法，平均数，离中趋势和编组频数分布的测度，不可避免地要丧失一定的精确性。这种精确性的丧失是否超过了在运算速度和呈现的方便上的所得，要由每个研究者为每一项研究项目重新决定。应用数据处理机减少了对作为概括性方法的编组资料及集中趋势测度的需要，但它们在呈现资料方面的优势依然存在。



时间数列的分析

由于本书所关心的是历史问题的计量分析，在其中时间维数总是存在的，但直到此刻“时间数列”问题才被明确提出可能有点令人感到意外，然而，重要的是认识到资料的历史特征并不意味着它们必然构成一个严格意义上的“时间数列”；我们只把“时间数列”的名称给予按照年月顺序排列的资料集。因而《末日裁判书》中有关教区的资料，以及关于商船的资料都不构成时间数列。它们是观察值的数列，首先是末日裁判的教区数列，其次是商船的数列，在时间上它们接近得使我们能够为每一资料矩阵中的所有观察值指定相同的日期；但这些项目却并非按时间排序的。

因而时间数列资料是资料矩阵中的一种特例，在其中资料是按照时间次序而不是其他次序排列的。在收集资料的过程中，可能会自然形成对资料按时间排序；如果我们以一系列时间的点，比如说每日、每月或者每年来收集某一变量的资料，那么我们的资料将按时间排序，而矩阵中的每一列都将成为一个时间数列。表 6.1 显示的就是这样一个资料矩阵；许多历史资料都属于这一类型：一系列人口普查年的人口数，每

表6.1　英国本土出口额　　1820—1850年

年份	百万英镑	年份	百万英镑
1820	36.4	1836	53.3
1821	36.7	1837	42.1
1822	37.0	1838	50.1
1823	35.4	1839	53.2
1824	38.4	1840	51.4
1825	38.9	1841	51.6
1826	31.5	1842	47.4
1827	37.2	1843	52.3
1828	36.8	1844	58.6
1829	35.8	1845	60.1
1830	38.3	1846	57.8
1831	37.2	1847	58.8
1832	36.5	1848	52.8
1833	39.7	1849	63.6
1834	41.6	1850	71.4
1835	47.4		

资料来源：B. R. 米切尔和P. 迪恩：«英国历史统计摘录»(B. R. Mitchell and P. Deane, Abstract of British Historical Statistics)，剑桥：剑桥大学出版社，1962年，第282页。

年的收获量，每月末的失业人数，以及其他许多历史资料。我们还可以选择另一种方式收集资料，这样矩阵中的个案（也就是矩阵的各行）不按时间排序，但是矩阵的列由按时间排序的信息组成；如表6.2所示，表的左边就是这样一个资料矩阵。或者把表中的各行按时间次序加以重新整理，或者建立一个频数分布，我们就可以从这个资料矩阵中建立时间数列资料。再则很多历史资料是以这种方式收集的；几乎所有关于个人的资料都将包含按时间排序的信息，如结婚和死亡日期，这些都可用于建立时间数列资料。

表6.2　从一资料矩阵求导时间数列

原始资料			
商船名称	建造地点	建造时间	吨位
占卜者	利物浦	1823	64
渴望	罗塞希提	1825	129
J.约克爵士	赤斯特	1822	62
马尔维纳	印威内斯	1824	39
创业	罗塞希提	1826	318
遮光	格陵诺克	1821	88
W.乔利夫	得普福	1826	235
旅游者	伯斯	1821	112
拉蒙娜	罗塞希提	1828	178
阿脱伍特	布拉克沃	1825	189
哈莱京	得普福	1826	185
伦敦市	得普福	1824	104
罗伐尔君主	得普福	1822	102
金星	罗塞希提	1821	112
索诃	布拉克沃	1823	292
贝尔法斯特	贝尔法斯特	1820	110
标枪	罗塞希提	1825	145
磁石	利梅豪斯	1826	166
简	北希尔兹	1826	12
利物浦伯爵	黑潭	1823	145
海王星	纽加塞耳	1824	42
詹姆斯·瓦特	格拉斯哥港	1821	291
耐久	巽特兰	1825	33
信使	罗塞希提	1826	103
庄严	格陵诺克	1821	143
易普威治	易普威治	1825	68
楼斗菜	得普福	1826	241
罗伐尔宪章	盖恩斯镇	1826	65
京斯敦	盖恩斯镇	1829	34

推导出的时间数列		
建造日期	商船数	吨位
1820	1	110
1821	5	746
1822	2	164
1823	3	501
1824	3	185
1825	5	564
1826	8	1325
1827	0	0
1828	1	178
1829	1	34

资料来源："所有在英国港口注册的蒸汽船的名称和种类"的统计表，《国会文件》第47卷，第545页。它们是1830年以前建造、1845年在伦敦港注册的蒸汽船。

由于时间数列资料只是排列资料矩阵的一种方式，我们已经讨论过的对资料矩阵加以概括和分类的种种技巧都可用于时间数列资料，只要资料的问题和性质宜于这样做。例如，我们能够计算表 6.1 中时间数列的平均数和标准差，其目的在于结合围绕平均数的离中趋势，发现 1820—1850 年之间英国本土出口额的平均值。与此类似，如果我们愿意，还可以计算时间数列资料的中位数和众数的值，并且运用所有我们已经讨论过的图解方法为时间数列资料制图。图 6.1 显示的就是表 6.1 中资料的曲线图。此外，我们可以应用若干不适用于按其他方式排列的资料的时间系列分析方法。

6.1 时间数列分析的对象及假设

表 6.1 给出 1820—1850 年产于大不列颠的商品的出口值。这是英国利用工业革命期间发展起来的新机器技术，非常迅速地发展起它的制造工业，并出口越来越多的工业制品的一个大好时期。如表 6.1 所示，它的本土出口值在 1820—1850 年[①]间几乎翻了一番。然而，这种增长并不是经常性的；贸易的不景气，其他国家的政治和经济事件，海外消费者爱好的转变，所有这些都影响了这一增长，所以一些年份的增长高于另一些年份。事实上，正如图 6.1 清楚地显示的那样，某些阶段在恢复向上发展之前甚至存在着一种下降。

如果分析这些年份英国出口的增长情况，则我们需要考

① 这些资料为货币价值，亦即没有根据物价水平的变动进行调整。这个问题下面还要较详细地讨论。

虑到可能会影响了这一增长的潜在因素，我们还需要某些方法，用以把一种因素的影响与其他因素的影响相区别。能够描述各种潜在影响而不对每个影响的重要性作出估计就没有多大意义。我们或许最感兴趣的不是出口的长期增长，而是年与年之间的短期波动，所以我们需能将短期变化从时间数列的长期变化中分离出来。因此我们需要能够将时间数列分解成各自对应于数列中的不同潜在影响的部分；而时间数列方法就是设计出来帮助我们做到这点的。这些方法并不能告诉我们对时间数列的影响是什么，因为那是一个历史学的问题。但是它们能帮助我们区别长时期影响和不过个别年份的短期影响；一旦我们区分开这些影响，就能运用我们的历史知识加以命名。

时间数列分析方法假定，可能有 3 种影响任何时间数列的类型。第 1 种影响在数列中引起长期的增长或下降，并被称为资料中的趋势。第 2 种影响为围绕长期趋势而出现的经常性波动。季节性波动就属于此列；在前工业化时期的英格兰，粮食价格在秋收之后总是降至最低点。另一个例子为不断变化的商业活动，在许多国家的经济中引起繁荣和衰退的交替，即所谓的“商业周期”。第 3 种影响是非经常性的，它在数列中引起短期的、不重复的波动。战争、瘟疫或者政府政策的改变都可能引起这类波动。所以，对时间数列进行统计分析在于把一个时间数列分解为若干与不同的潜在影响相对应的部分，如长期影响、短期影响、经常性的波动及非经常性的波动，等等。我们运用时间数列分析方法就能够将每一项以及全部影响分离出来。

时间数列分析的假设是：一个时间数列由上述 3 种类型

影响的结果所构成。在用时间数列方法进行分析时，我们正是接受这一假设。因此，我们必须十分小心，不要使这一假设与我们作为历史学家对某一特殊资料集的了解发生冲突。如果我们作为历史学家，不相信有一个经常性的周期影响对某一特殊数列起着作用，就不该运用一种假设这种影响存在着的分析方法。在本章的以后部分我们还要谈到这个问题，现在先让我们讨论对图 6.1 的时间数列的分析。

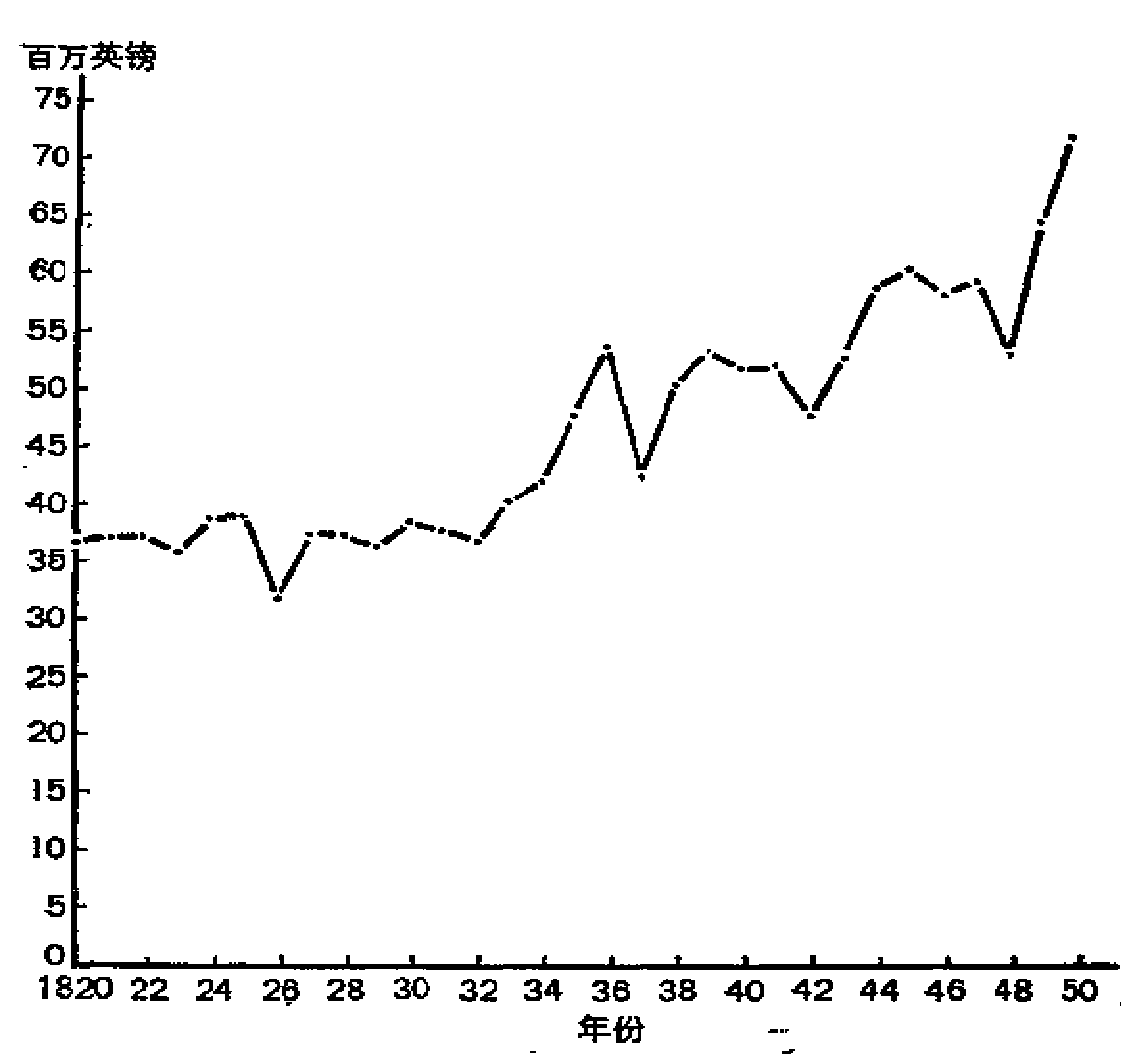


图 6.1　1820—1850 年英国本土出口
资料来源：表 6.1。

6.2 增长率

图6.1清楚地表明1820—1850年之间英国本土出口呈上升趋势。然而，若我们要想详细地分析这一增长，还要了解在整个时期以及部分时期这一增长快到什么程度。例如，我们会有兴趣了解，这一增长究竟是否在初期比末期更快这样的问题。另外，例如若我们想要比较出口增长与进口增长，我们就需要某种对增长的测度，用以比较两个数列。

从表6.1可以看出，1820—1850年出口几乎翻了一番。很明显这是一个很大的增长，但用“31年间几乎翻了一番”这种表达形式，这样就难与其他较短或较长时期的变化相比较。如果我们能将这一增长表示为每年的平均增长，这种比较就方便了。似乎计算这样一个年平均增长非常容易，只要把1820年与1850年的出口额之差除以这一期间的年数即可；我们将得到

$$\frac{71.4-36.4}{30}=1.167$$

这样我们就可以说英国出口每年增加的平均数为1.167（百万）英镑。这是不错的，但它对我们与其他数列作比较方面帮助不大，因为我们不会知道增长起始的那个基数。以1（百万）英镑为起点每年增加1.167（百万）英镑较之以100（百万）英镑为起点每年增加相同数量要引人注目得多，可是上述的增加平均数不能区分这两种情况。如果我们不顾原来的单位（在本例中是英镑），要把我们的出口值与以英担为单位的茶叶进行比较，那也是有用的。

这两种必要条件，即需要考虑基数以及需要有一个不顾原来单位的测度，提示只有基于百分比的测度才会是合适的。这里还有一个更进一步的要求，即这一测度应是累加的，把每年的增长表达为以前一年值的百分比；事实上它应以复利率，而不是以单利率计算。所以满足上述所有要求的增长率为百分比增长率，其计算公式如下

$$r = \left(\sqrt[m]{\left(\frac{X_N}{X_T}\right)} - 1\right)100$$

这里 r 等于所求的增长率，X_N 为末期值，X_T 为初期值，m 是初期与末期之间年份之差。

应用对数可以大大简化增长率的计算。以表 6.1 中的资料为例，为计算 1820 年与 1850 年的增长率，我们查出对数 $X_N(\log 71.4 = 1.8537)$ 和 $X_T(\log 36.4 = 1.5611)$。两者相减，$\log X_N - \log X_T = 0.2926$。为开 30 次方需除以 $m = 30$，得到 0.0098。0.0098 的反对数为 1.023，减 1 再乘以 100，我们就得到每年平均百分比增长率为 2.3%。

（应用印就的增长率表也可以查出增长率，而不劳计算。还应指出，尽管在上述的例子里我们计算的是每年的增长率，运用同样的方法我们可以计算任何一段时期的增长率。）

增长率是一种很有价值并被广泛应用的描述时间数列资料的方法，但使用它们时要小心，当数列中有显著的波动时尤为如此。在这种情况下，选择增长率据以计算的起始和终止年份极为重要。我们可以从表 6.1 中的资料计算出其他一些增长率来说明这点；表 6.3 显示了这些增长率，图 6.2 则将它们绘制在半对数尺度图上。由于增长率是对按不变比例（百分数）的增加率的测度（如每年百分之 2.3），所以半对数图是

合适的。如第四章所述，运用对数垂直尺度便给出这样一个图表，它以一条直线（但不是水平线）代表按不变比例的变化，这条线越陡就说明变化越快；因而在图 6.2 中，1820—1850年的连结线和 1826—1850 年的连结线在斜率上的差恰恰表示后一时期的增长较快。在一张线性尺度图中，按经常比例的变化由一条曲线表示（如图 6.1），这种曲线较难绘制，并在曲线之间进行比较不大容易。

表 6.3 中所有的增长率都是正确的，它们给出不同起迄年份中的每年百分比增长率。可是，作为整个时间数列的一个测度，它们各有严重的缺陷；我们从 1820—1848 年英国出口增长得到的印象与从 1826—1850 年之间增长得出的印象大不相同，尽管我们计量的时期并无很大的不同。

表 6.3　表 6.1 中资料的增长率

起迄年份	时期长度	每年百分比增长率
1820—1850	30年	2.3
1820—1848	28年	1.4
1826—1850	24年	3.5
1823—1847	24年	2.1

如果我们研究一下图 6.2，其中表 6.3 中所用的起迄年份都已用粗线连结起来，表 6.3 中产生增长率之间差别的原因就清楚了。我们见到，在选择 1820—1848 年为起迄年份时，我们是从 1820 年的高点至 1848 年的低点进行测度的，而在 1826—1850 年这段时间我们所做的恰恰相反。甚至当我们取 1820—1850 年为起迄年份时，我们仍可以看到连结这两个年份的线超越了几乎所有其他的资料点。只是在取1823—1847 年为起迄年份时，我们似乎才选择了据以计算增长率的

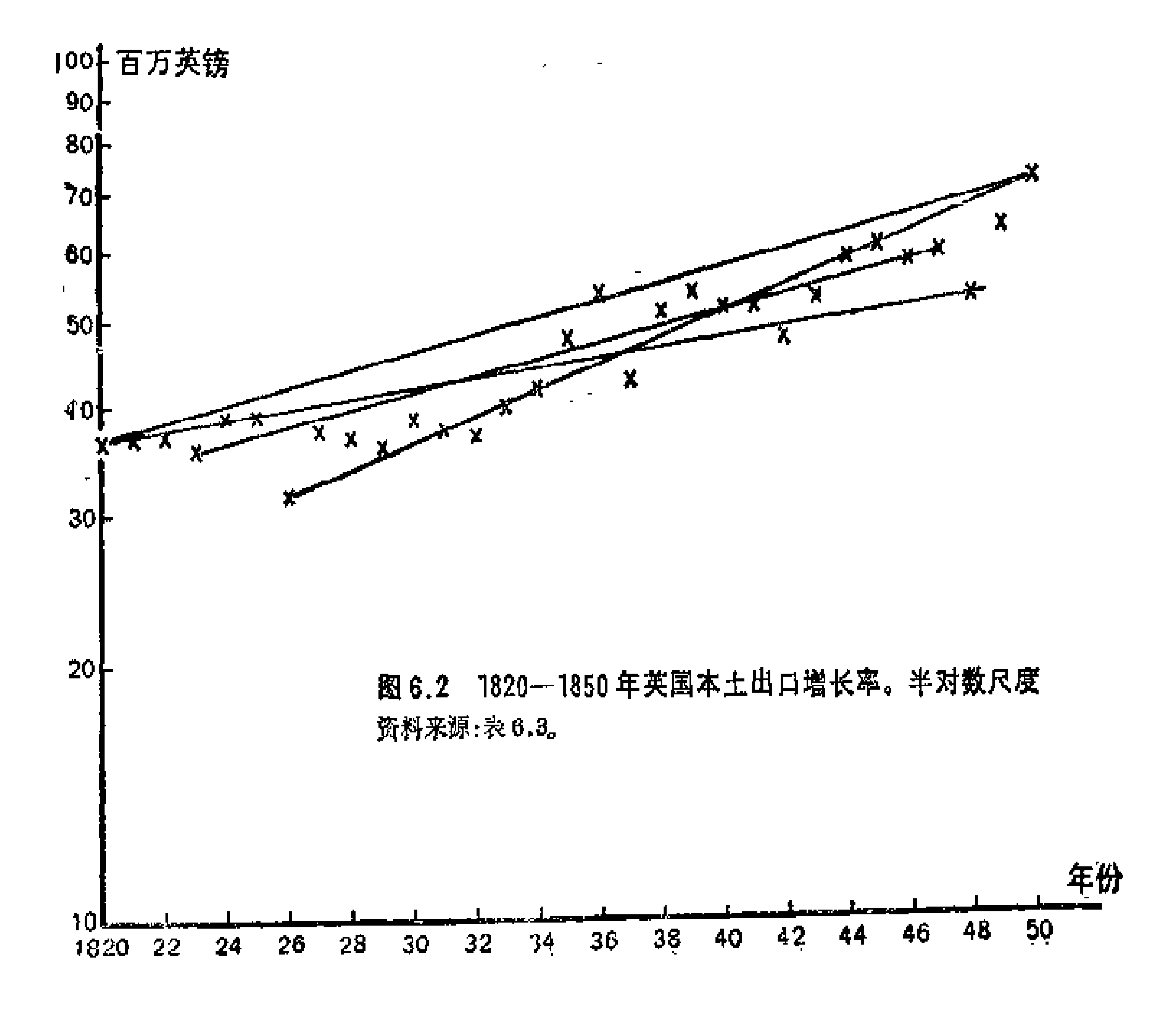


图 6.2　1820—1850 年英国本土出口增长率。半对数尺度

资料来源：表 6.3。

具有相当典型性的年份。

这绝不仅是个统计问题。为了选择用于这类计算的起迄年份，我们需要极其小心以免选择非典型性的年份，因为这样做会严重影响我们的结果。例如，有关工业革命时期英国工人生活水准变化的长期的历史学争论大部分与选择合适的起迄年份有关。如果选择了这些年份，生活费用会明显上升；如果选择了那些年份，生活费用则会下降。

6.3 趋势

如果我们再看图 6.2，选择适当的起迄年份的部分困难，似与我们必须选择两个而且只有两个起迄年份来描述整个数列的增长这一事实有关。因此，其他年份的资料不被计算在内。如果我们要去找到一种对整个数列增长的测度，去找到一种使我们能应用整个数列的测度，则显然更为合理有用。事实上，我们所要描述的是资料中的长期趋势，亦即在时间数列分析中假设的第一种影响对时间数列所发生的作用。

我们将从英国出口每年按一不变的绝对量增长这一假设开始，然后用数列中的所有资料对这个绝对量作出估计。一个按不变的绝对量变化的数列可以用一条直线在自然尺度图上表示出来，因此我们将用图 6.1 所示的资料。在本章的以后部分中我们将会考虑一种当我们不是把一个数列看作每年按不变的绝对量，而是按不变的比例增长时为合适的方法；这样的数列以及在这种数列中的长期趋势可以用一条曲线在自然尺度图或用一条直线在半对数图（如图 6.2）上表示出来。对于一个资料数列中的长期趋势，我们有时很难决定究竟是

用一条直线("线性")形式还是用一条曲线("曲线性")的形式来表示为最好，因此我们必须事先懂得适用于每一种情况的方法。

在上一节里，我们计算出1820—1850年每年出口的平均绝对增长为1.167(百万)英镑。在计算中仅用了在1820年和1850年的出口值资料，不顾介于两者之间的其他年份。从图6.3可以看出，这种计算等同于在一自然尺度图上画一条连结1820年和1850年数据点的线，并测度每一年沿水平轴向前运动时这一连线沿垂直轴上升所呈现的距离。然而，正如在上一节所看到的那样，仅使用1820年和1850年的资料可能会导致误解，因为这两个年份可能是非典型性的。最理想的是，为了充分考虑到每一年份的资料，我们需要找到一条穿过所有资料点的直线。然而，从对图6.3的观察中可以明显看出，我们不可能找到这样一条连结图中所有资料点的直线。因此，作为一个较次的最好办法，我们试图找到一条尽可能接近所有资料点的直线，从某种意义上说这条线是所有可能穿过图中资料点的直线的平均数。一些资料点将处于这条线上，其他数据点或高于或低于它。当然，我们可以根据自己对这条直线应处方位的判断，试在图中画出这条线，但是我们的判断很可能会出错，并引起别人的争议。因此，我们需要根据某种理论来计算这条线，这种理论被认为能对图中所有资料点作出最佳拟合。

可以显示出，这条最佳拟合直线是通过"最小平方法"计算出的。图6.4阐明了这一方法的逻辑。我们试选择这样一条穿过图中资料点的直线(即图6.4中的线B)，然后从每一点向这条线引垂直线(如我们对1835—1845年所做的那样)，计

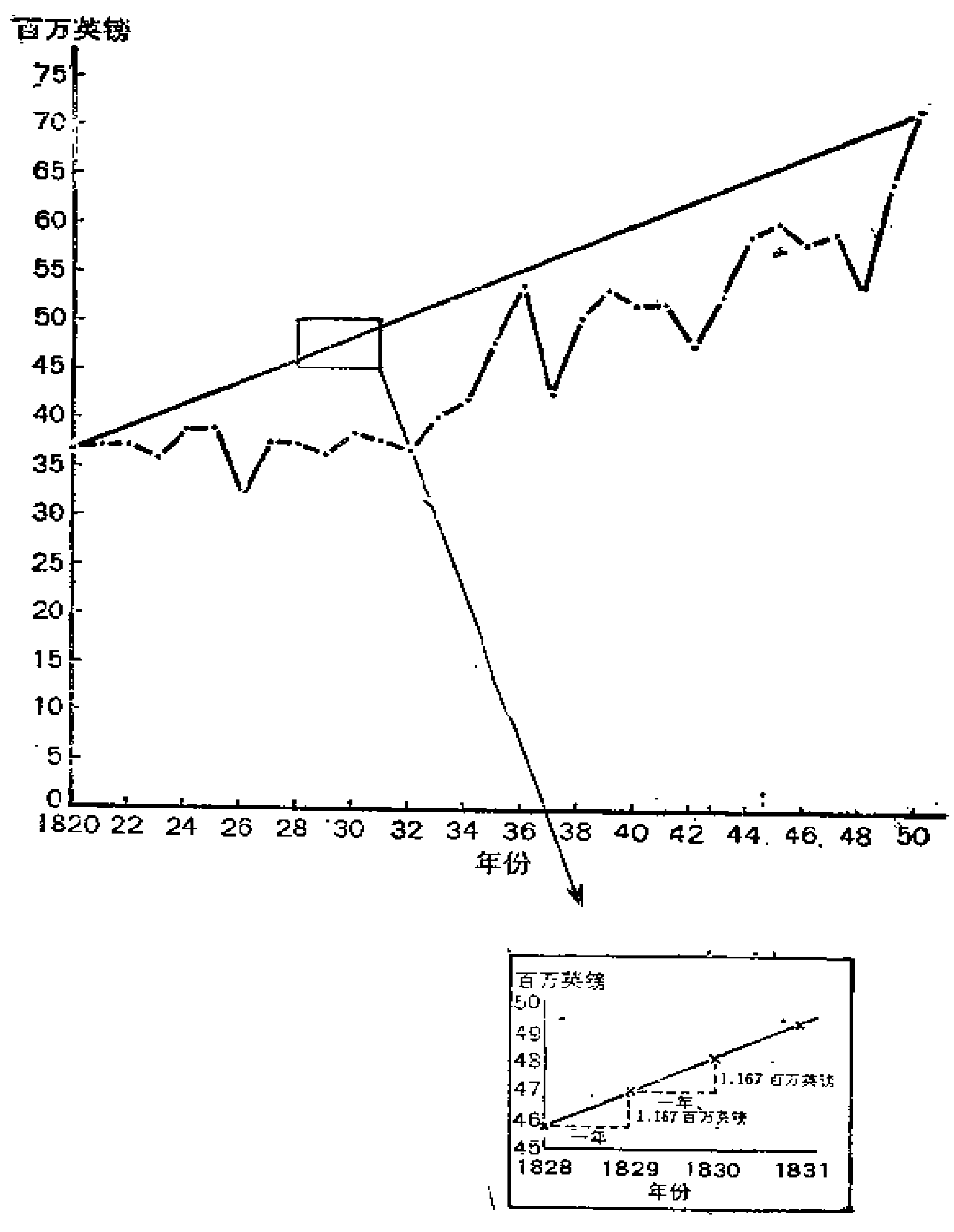


图 6.3　1820—1850年英国本土出口。
1820—1850 年的平均年增长

资料来源:图 6.1。

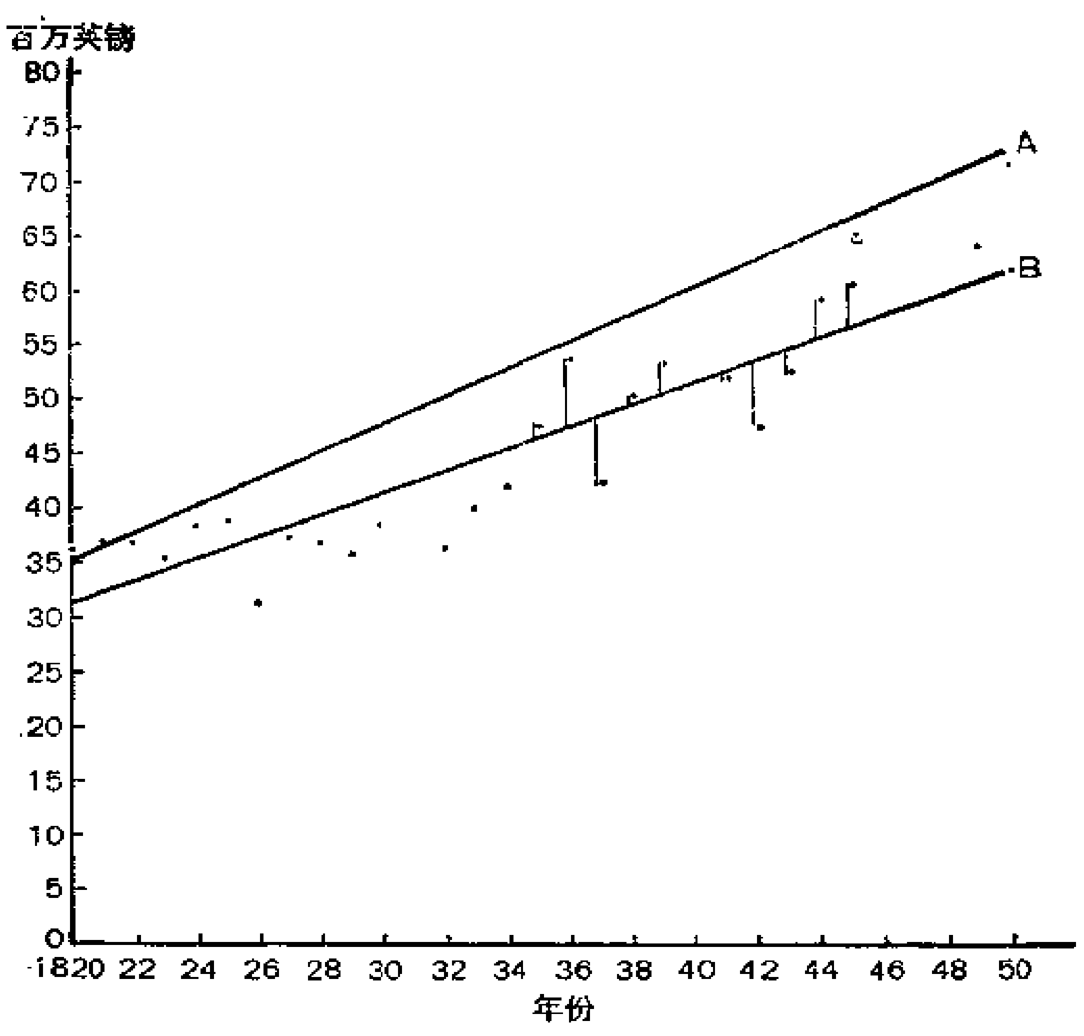


图 6.4 英国本土出口，1820—1850 年。趋势

资料来源：表 6.1。

量这些连线间的距离，对每一个值平方后再相加，这时可以看出从我们选择的这条线得出的结果（所有点与此线偏离的平方和）是小于图中画的任何其他线可能得出的结果。用这种方法找到的这条直线对图中所有点提供最佳的拟合，因此它也就是用以描述资料里长期趋势的最合适的直线。

我们可以通过试误法，在图上画线，测度偏差，计算平方和等等，发现这一最佳拟合线，但很明显这将是一个很麻烦的

过程。我们代之以应用两个公式，两者一起可使我们绘制这条直线时满足偏离平方之和应尽可能小的这个条件。为了理解这一点，我们必须考虑如何绘制如图 6.4 中的线 B 那样的线。线 B，或者任何像图中所画的线 A 那样的直线，都有两个重要特征。其一，如此线一直沿图中垂直轴延伸，它必然与那个轴在一特定点相交。其二，线的倾斜与图的水平轴相对。

为了在图中画出一条特定的直线，我们需要知道关于它的两点：第一，它与竖轴在哪里相交；第二，此线相对于水平轴的斜率是多少。第一点，称为截距，仅是垂直轴上单位的一个数目。如图 6.4 中的线 A，与垂直轴相交于代表 35（百万）英镑的点上。第二点稍微复杂一些。以截距为起点沿水平轴画线时我们需知道此线向上或者向下的幅度。再以线 A 为例，我们看到 1820 年表示的值是 35（百万）英镑，1822 年为 37.5（百万）英镑，1824 年为 40（百万）英镑，以此类推。换句话说，沿水平轴每前进一年，我们都要在垂直轴上加上 1.25（百万）英镑。10 年以后我们应已累加了 12.5（百万）英镑，20 年以后累加了 25（百万）英镑，而我们可以知道线 A 实际上是一条追踪这些值的直线。

因此，为了能画出最能拟合我们的资料的线 B，我们需要知道两点：截距和斜率，这也是依据最小平方法的两个公式给我们的两项信息。这两个公式是

$$\text{截距，}\quad a=\frac{\sum Y-b\sum X}{N}$$

$$\text{斜率：}\quad b=\frac{N\sum XY-(\sum X)(\sum Y)}{N\sum X^2-(\sum X)^2}$$

表6.4 根据表6.1资料，线性趋势的计算(长方法)

时间单位以1820年为基础				
年份	资料(Y)	(X)	X^2	XY
1820	36.4	0	0	0
1	36.7	1	1	36.7
2	37.0	2	4	74.0
3	35.4	3	9	106.2
4	38.4	4	16	153.6
5	38.9	5	25	194.5
6	31.5	6	36	189.0
7	37.2	7	49	260.4
8	36.8	8	64	294.4
9	35.8	9	81	322.2
30	38.3	10	100	383.0
1	37.2	11	121	409.2
2	36.5	12	144	438.0
3	39.7	13	169	516.1
4	41.6	14	196	582.4
5	47.4	15	225	711.0
6	53.3	16	256	852.8
7	42.1	17	289	715.7
8	50.1	18	324	901.8
9	53.2	19	361	1010.8
40	51.4	20	400	1028.0
1	51.6	21	441	1083.6
2	47.4	22	484	1042.8
3	52.3	23	529	1202.9
4	58.6	24	576	1406.4
5	60.1	25	625	1502.5
6	57.8	26	676	1502.8
7	58.8	27	729	1587.6
8	52.8	28	784	1478.4
9	63.6	29	841	1844.4
50	71.4	30	900	2142.0

$\Sigma Y=1429.3 \quad \Sigma X=465 \quad \Sigma X^2=9455 \quad \Sigma XY=23973.2$

$$b=\frac{743169.2-664624.5}{293105-216225}=\frac{78544.7}{76880.0}=1.02$$

$$a=\frac{1429.3-1.02(465)}{31}=\frac{1429.3-474.3}{31}=30.81$$

$$\therefore Y_T=30.81+1.02X_T$$

在这些公式中，像通常一样，N是所有值的数目。当我们计算一个时间数列的趋势时，X为从时间数列开始以来的年份的向量，而Y为资料值的向量。表 6.1 中资料的 2 个向量X和Y，在表 6.4 中被表示为第 3 和第 2 列。表 6.4 还显示了求解 a 和 b 公式所需其它数值的计算方法。

如表 6.4 所示，计算了几项和以及几项平方和之后，我们就可以先计算 b，然后 a。我们发现，对于这些资料，直线的截距为 30.81，斜率为 1.02。这意味着此线与垂直轴相交于 30.81(百万)英镑这一值点，每年沿水平轴上升 1.02（百万）英镑。至此，我们知道 1820 年此线经过代表 30.81(百万)英镑的值点，1821 年此线经过代表 31.83（百万）英镑的值点，1830 年此线经过 30.81 + 10(1.02) = 41.01（百万）英镑的点，依此类推，此线即是线 B，而且我们说根据最小平方法它是对表 6.1 中的资料的拟合；它代表时间数列中的“线性趋势”。

通过应用一个简单的公式，我们能计算出线 B 将经过的所有值点

$$Y = a + bX$$

这里对 a、b、X 和 Y 的定义与在求 a、b 的最小平方法公式中的定义一致。此公式为求直线的一般公式，根据 a 和 b 的变化可以描绘任何特定直线。例如，我们可以用等式

$$Y = 35 + 1.25X$$

在图 6.4 中画出线 A，而用等式

$$Y = 30.81 + 1.02X$$

画出线 B。与此类似，如果我们记得曾经计算过 1820—1850 年之间每年绝对增长的平均值为1.167(百万)英镑，我们便可

表 6.5　根据表 6.1 资料，线性趋势的计算（简便方法）

年份	资料值 (Y)	时间单位以1820年为基础	时间单位 (X)	X^2	XY	趋势值
1820	36.4	0	−15	225	−546.0	30.81
21	36.7	1	−14	196	−513.8	31.83
22	37.0	2	−13	169	−481.0	32.85
23	35.4	3	−12	144	−424.8	33.87
24	38.4	4	−11	121	−422.4	34.89
25	38.9	5	−10	100	−389.0	35.91
26	31.5	6	−9	81	−283.5	36.93
27	37.2	7	−8	64	−297.6	37.95
28	36.8	8	−7	49	−257.6	38.97
29	35.8	9	−6	36	−214.8	39.99
30	38.3	10	−5	25	−191.5	41.01
1831	37.2	11	−4	16	−148.8	42.03
32	36.5	12	−3	9	−109.5	43.05
33	39.7	13	−2	4	−79.4	44.07
34	41.6	14	−1	1	−41.6	45.09
35	47.4	15	0	0	0	46.11
36	53.3	16	+1	1	+53.3	47.13
37	42.1	17	+2	4	+84.2	48.15
38	50.1	18	+3	9	+150.3	49.17
39	53.2	19	+4	16	+212.8	50.19
40	51.4	20	+5	25	+257.0	51.21
1841	51.6	21	+6	36	+309.6	52.23
42	47.4	22	+7	49	+331.8	53.25
43	52.3	23	+8	64	+418.4	54.27
44	58.6	24	+9	81	+527.4	55.29
45	60.1	25	+10	100	+601.0	56.31
46	57.8	26	+11	121	+635.8	57.33
47	58.8	27	+12	144	+705.6	58.35
48	52.8	28	+13	169	+686.4	59.37
49	63.6	29	+14	196	+890.4	60.39
1850	71.4	30	+15	225	+1071.0	61.41
	1429.3			2480	+2533.7	

$$a=\frac{\Sigma Y}{N}=\frac{1429.3}{31}=46.11$$

$$b=\frac{\Sigma XY}{\Sigma X^2}=\frac{2533.7}{2480}=1.02$$

$$\therefore Y_T=30.81+1.02X_T$$

以说图 6.3 中连结这两个年份的直线可用下边的等式来描绘

$$Y = 36.4 + 1.167X$$

然而，只有图 6.4 中的线 B 才具备对所有资料点进行最佳拟合的所需特性。

表 6.5 所示的计算方法可以减轻计算线性趋势时的繁重计算工作。如果我们不像在表 6.4 中那样从时间数列的开始测度作为时间单位的 X，而是从时间数列的中间开始测度，因此 $\sum X = 0$。若这样做，最小平方法的两个等式就变成简单得多的形式

$$a = \frac{\Sigma Y}{N}$$

及

$$b = \frac{\Sigma XY}{\Sigma X^2}$$

而我们便像表 6.4 那样计算线性趋势。

为了依据线性趋势计算增长率，仅需在线性趋势线上取两个值点，这两个值点对应于我们想要测度增长的年份。例如，从线 B 的公式 $Y = 30.81 + 1.02X$ 中我们知道 1820 年的值是 30.81，1850 年的值是 61.41；计算这一起迄年份之间的增长率，我们得到的年增长率为 2.3%。（在表 6.5 中我们计算出 $a = 46.11$；这是因为取了 1835 年为时间单位中点的缘故，因此当我们计算 1835 年的趋势值时，$X_{1835} = 0$，Y_{1835} 就等于 46.11。）用同样的方法我们可以计算趋势线上任意两点间的增长。

在上述例子中，1820—1850 年线性趋势的增长率与根据资料的起迄年份计算的增长率是一致的；这个一致纯属巧合，

从理论上看这两种计算增长率的方法完全是分立的，尽管在上面的例子中两者的结果相同。总的说来，由于线性趋势考虑了时间数列中的所有单个值，根据线性趋势计算增长率要比较可取得多。

一旦对资料中的线性趋势作出估计，我们就可从原始资料值中减去趋势值，如表 6.9 中第 2 和第 3 列所示。其结果为一个由原始数列中的波动所组成的时间数列，因此我们就可不受资料包含长期趋势种种复杂因素的影响，继续对这些波动进行分析。与趋势偏离的时间数列用图 6.5 表示。

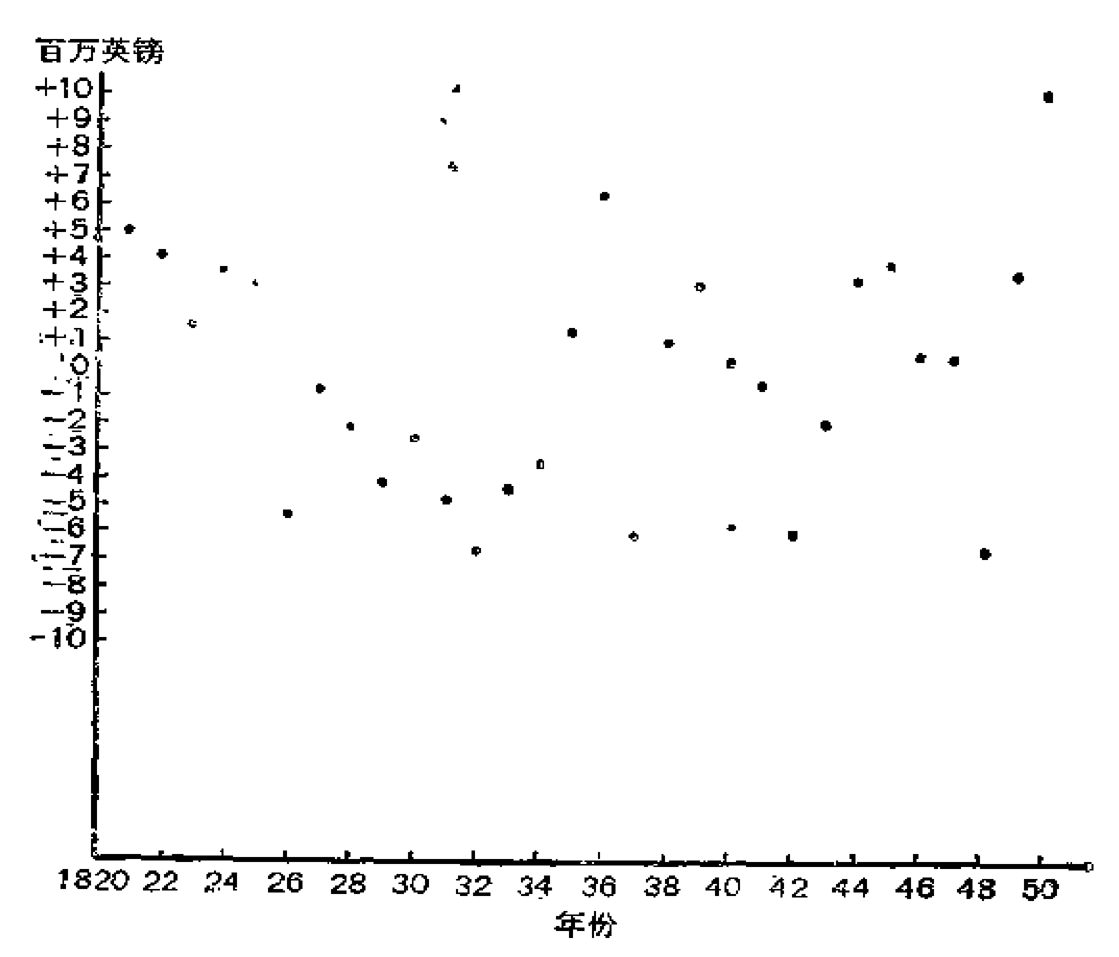


图 6.5　1820—1850 年英国出口趋势偏离

资料来源：表 6.1。

迄今我们已集中讨论了线性趋势作为一种概括一个时间数列的方法的优点，以及集中注意于环绕趋势的波动的优点。但这种方法是设计来以一条直线拟合于资料。而很多历史时间数列并不见得是直线性的，因而有其缺点。例如如果我们比较图 6.7 和显示 1770—1800 年英国原棉进口增长的图 4.7，就不难发现一个为适当的线性形式的数列与一个为曲线形式的数列之间的差别。当数列是非直线性时，很明显不适宜用我们刚刚讲过的以一条直线来表述资料的方法，因为这样一条直线可能使人对趋势产生错误的印象；在此用一条曲线来表示这一趋势更为合适。尽管常常可以做到这一点，但它所涉及的数学内容要比计算一个线性趋势所涉及的数学内容要复杂得多。因此，经常采用的一个替代方法是将数列转变成对数形式。这样做非常合理，正像我们在上节中在看增长率时所见到的那样，我们可以假定数列每年按一个相同比例数量变化。如果原棉进口的资料标绘在一个半对数图上，如图 4.9 所示，我们就可以看出各个值呈现出一条非常接近于直线的形式，因此在这里用原始资料的对数值计算线性趋势非常适当。完成这项工作之后（如表 6.6），趋势值就可以被标绘在一张半对数图上（如图 6.6 所示），得出一条直线，或则在原来的图上经过反对数的转换之后，将趋势值给出一条曲线，如图 6.7 所示。这一方法的更进一步的优点是，增长率可以立即从经过计算的趋势等式中显现出来。这个等式为 $Y=ab^X$。其中 b 等于 1 加上每年平均增长率。因此，可以马上看出原棉进口增长率为每年 9.3%。实际上，这一优点很大，因此用对数计算线性趋势常比用原始资料计算线性趋势更为可取，就是当原始资料近似于一条直线时也是如此。

表 6.6　根据 1770—1800 年英国进口原棉资料

运用对数计算的线性趋势

年份	资料 (Y)	logY	X	X^2	XlogY	log 趋势值	趋势值
1770	3612	3.5577	−15	225	−53.3655	3.5369	3443
1	2547	3.4060	−14	196	−47.6840	3.5758	3765
2	5307	3.7249	−13	169	−48.4237	3.6147	4118
3	2906	3.4633	−12	144	−41.5596	3.6536	4504
4	5707	3.7564	−11	121	−41.3204	3.6925	4926
5	6694	3.8257	−10	100	−38.2570	3.7314	5388
6	6216	3.7935	−9	81	−34.1415	3.7703	5892
7	7037	3.8474	−8	64	−30.7792	3.8092	6445
8	6569	3.8175	−7	49	−26.7225	3.8481	7049
9	5861	3.7680	−6	36	−22.6080	3.8870	7709
1780	6877	3.8374	−5	25	−19.1870	3.9259	8431
1	5199	3.7160	−4	16	−14.8640	3.9648	9221
2	11828	4.0730	−3	9	−12.2190	4.0037	10090
3	9736	3.9884	−2	4	−7.9768	4.0426	11040
4	11482	4.0599	−1	1	−4.0599	4.0815	12060
5	18400	4.2648	0	0	0	4.1204	13190
6	19475	4.2896	1	1	4.2896	4.1593	14430
7	23250	4.3664	2	4	8.7328	4.1982	15790
8	20467	4.3111	3	9	12.9333	4.2371	17260
9	32576	4.5130	4	16	18.0520	4.2760	18880
1790	31448	4.4976	5	25	22.4880	4.3149	20650
1	28707	4.4581	6	36	26.7486	4.3538	22580
2	34907	4.5429	7	49	31.8003	4.3929	24710
3	19041	4.2797	8	64	34.2376	4.4316	27020
4	24359	4.3867	9	81	39.4803	4.4705	29540
5	26401	4.4216	10	100	44.2160	4.5094	32310
6	32126	4.5069	11	121	49.5759	4.5483	35340
7	23354	4.3683	12	144	52.4196	4.5872	38660
8	31881	4.5035	13	169	58.5455	4.6261	42280
9	43379	4.6373	14	196	64.9222	4.6650	46240
1800	56011	4.7483	15	225	71.2245	4.7039	50580
		127.7309		2480	96.4981		

$$\log a=\frac{\Sigma \log Y}{N}=\frac{127.7309}{31}=4.1204$$

$$\log b=\frac{\Sigma X\log Y}{\Sigma X^2}=\frac{96.4981}{2480}=0.0389$$

$\log Y=4.1204+0.0389X$　　取反对数　$Y=(13190)(1.093)^X$

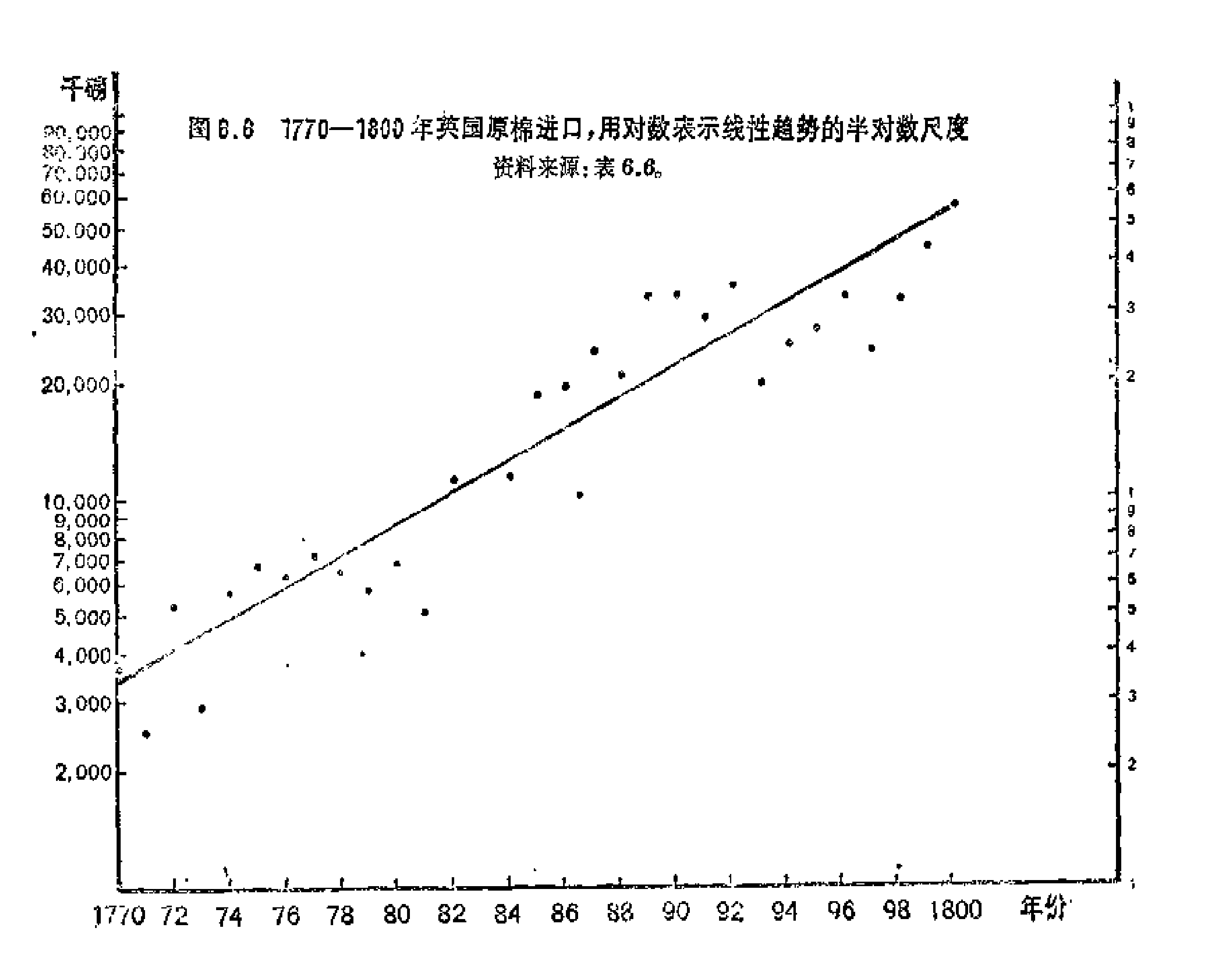


图 6.6　1770—1800 年英国原棉进口，用对数表示线性趋势的半对数尺度

资料来源：表 6.6。

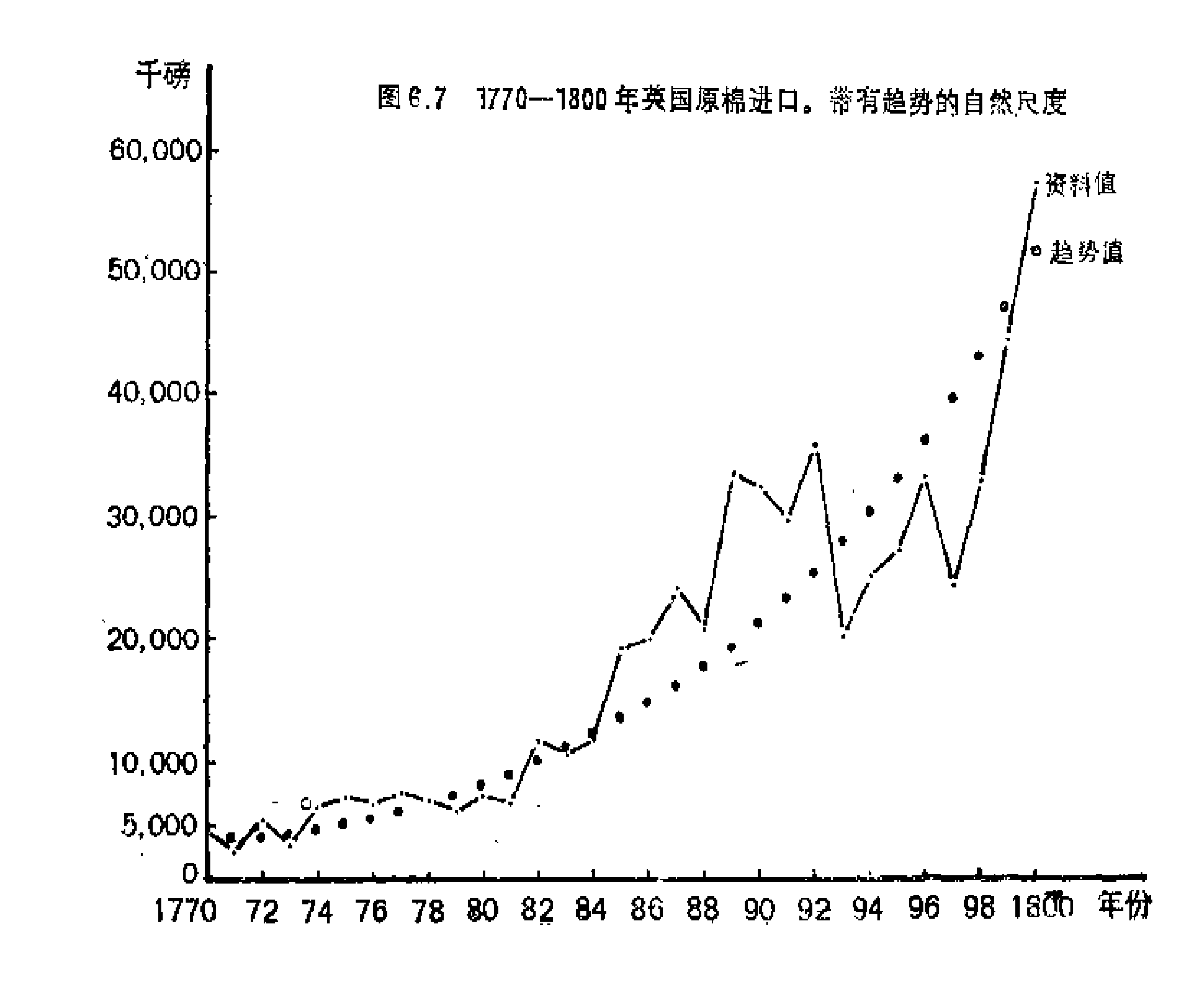


图6.7　1770—1800年英国原棉进口。带有趋势的自然尺度

6.4 时间数列中的经常性波动

上一节里我们集中谈了运用线性趋势计算增长率的优点。另外，如我们见到的，线性趋势代表了时间数列中长期因素的影响。例如，在表 6.1 资料的例子里，我们可以认为线性趋势指出了英国制造能力的长期增长及外国对英国产品需求的增长。除了这一出口增长的长期趋势以外，图 6.5 清楚地表明围绕着这一趋势存在着一些显著的波动，现在我们就可以着手讨论如何分析这些波动。

时间数列分析方法假定有 3 类可能的波动，其中 2 种是经常性的，1 种是非经常性的。第一类有经常性的波动被称为季节性波动，它包括气候，一星期或一年里工作和闲暇型式，以及其他经常性的每星期、每个月或每年的所发生事件等因素所引起的波动。在前工业化社会里，这类波动，特别是由气候引起的波动，对工作型式和生活的很多方面发生相当大的影响：冬季难于旅行，夏天磨坊缺乏水源动力，冬季和收获之前食品价格上涨。甚至在现代化社会里，每年像圣诞节和复活节这样的特定节日也会影响工作型式，并且食品价格仍然在各季节里变动。因此，由所收集的间隔不到一年的资料所组成的任何时间数列，都有可能受到季节性波动的影响。像表 6.1 的时间数列，由于资料只是按年收集的，当然不可能受到季节变化的影响。

为了说明从时间数列中分离季节性波动的方法，我们因此要使用另一集资料，它由 1713—1718 年温彻斯特学院购买小麦的价格所组成。这些价格是由贝弗里奇勋爵作为他对英

表 6.7　计算 1713—1718 年温彻斯特学院小麦价格的季节性波动

时　期		小麦价格（布夸特小麦价格）	趋势值[1]	趋势值的偏离	季节成分[2]	非趋势、非季节性数列	非季节性数列
1713	1st 季度	42.67	46.71	−4.04	−0.09	−3.95	42.76
	2nd季度	56.88	45.86	11.02	1.55	9.47	47.41
	3rd季度	49.78	45.01	4.77	0.73	4.04	49.05
	4th 季度	46.21	44.16	2.05	−2.19	4.21	48.40
1714	1st 季度	32.00	43.31	−11.31	−0.09	−11.22	32.09
	2nd季度	32.00	42.46	−10.46	1.55	−12.01	30.45
	3rd季度	32.00	41.61	−9.61	0.73	−10.34	31.27
	4th 季度	28.44	40.76	−12.32	−2.19	−10.13	30.63
1715	1st 季度	46.21	39.91	6.30	−0.09	6.39	46.30
	2nd季度	49.78	39.06	10.72	1.55	9.17	48.23
	3rd季度	42.67	38.21	4.46	0.73	3.73	41.94
	4th 季度	35.56	37.36	−1.80	−2.19	0.39	37.75
1716	1st 季度	39.10	36.51	2.59	−0.09	2.68	39.19
	2nd季度	39.10	35.66	3.44	1.55	1.89	37.55
	3rd季度	40.29	34.81	5.48	0.73	4.75	39.56
	4th 季度	33.77	33.96	−0.19	−2.19	2.00	35.96
1717	1st 季度	43.84	33.11	10.73	−0.09	10.82	43.93
	2nd季度	32.00	32.26	−0.26	1.55	−1.81	30.45
	3rd季度	32.00	31.41	0.59	0.73	−0.14	31.27
	4th季度	32.00	30.56	1.44	−2.19	3.63	34.19
1718	1st 季度	24.89	29.71	−4.82	−0.09	−4.73	24.98
	2nd季度	23.70	28.86	−5.16	1.55	−6.71	22.15
	3rd季度	26.67	28.01	−1.34	0.73	−2.07	25.94
	4th 季度	24.89	27.16	−2.27	−2.19	−0.08	27.08

1. 趋势值是根据估计的线性趋势等式，价格＝36.93－0.85时间，用简捷方法算出的。

2. 季节性成分是通过取每年第一季度的趋势偏差平均数，及第二季度的趋势离差平均数，等等计算出的。得出－0.08，1.56，0.74，－2.17；将这些值合计得到0.05，但不言而喻季节变量对全年的影响应是中性的或为零。所以，我们用每一个案(0.05/4）的近似值去调整季节平均数，得到－0.09，1.55，0.73，－2.19，其和为零，用这些值作为季节变量的估计值。

资料来源：贝弗里奇：《英格兰的价格与工资》(Beveridge, Price and Wages in England)，第 1 卷，第82页。

国价格和工资的伟大研究的一部分而收集的，它们是研究生活水准的重要凭据。我们需要从数列中排除季节性影响，主要使我们能研究特定年份中可能引起饥荒或供应过剩的短期波动，另一方面还使我们能研究不受每年经常性变化影响的价格的长期趋势。

为了分离出季节因素，我们首先必须估计趋势值，因为从表 6.7 和图 6.8 可以看出这一数列有一个下降的趋势。如果不消除这一趋势的影响，它可能会左右我们对季节变量的估计。因此，我们估计线性趋势，计算出数列与趋势值的偏差，给出在表 6.7 的第 3 列中所示的数列。为了计算出经常的季节性成分，我们要取每年第一季度的所有值，计算它们的算术平均数，并对第二、第三、和第四季度的值也进行这一计算工作，在表 6.7 的第 4 列给出这些值。这些值代表每一季度的平均上涨或下降数列，这也就是我们所说的季节性成分。把这些季节值从趋势偏差中减去所得的残差（第 5 列）代表长期趋势和季节性成分以外的因素的影响。我们也可以（如第 6 列）把季节性成分从原始数列中减去，得到的再一个数列包含长期趋势和残差波动，但排除了季节性影响。

如果我们的资料是按星期或月收集的，我们将遵循完全相同的步骤，找出一年中对应的星期和月的偏差平均数。应该指出，还有其他分离季节性变量的方法，参考书目中开列的统计学著作对此有更为详细的描述。

从资料中排除了季节性波动后，还剩下两类波动，一类是经常性波动，另一类是非经常性波动。第一类经常性波动正规地被称为周期波动。它与季节性波动的区别在于周期波动发生的间隔往往长于一年。这些波动最常见的形式就是被经

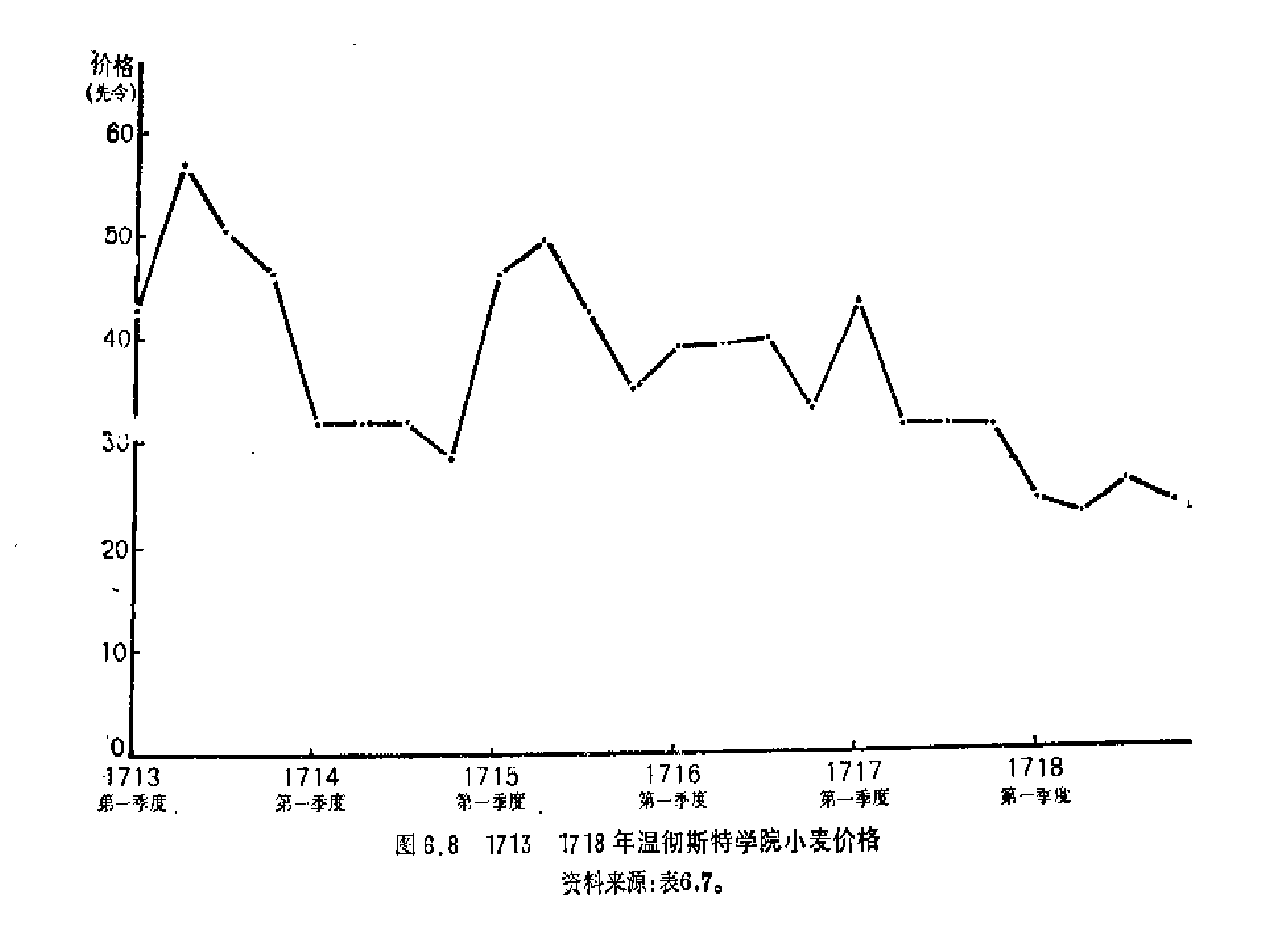


图6.8 1713—1718年温彻斯特学院小麦价格

资料来源:表6.7。

济史学家用来描述经济活动中经常性或半经常性变化的商业或贸易周期，这些变化在19世纪最为显著。商业通常每隔7—10年就经历一个从萧条到繁荣，再转入萧条的经济周期，虽则有些经济史学家试图把那些持续达几世纪的“长波”区别开来。

如前所述，时间数列分析方法假定时间数列中可能存在着周期性波动，因此它提供了分离这些波动影响的手段。当然，历史学家是否运用这些分离周期性波动的手段完全取决于他是否认为在他的历史数列中存在着这类周期。这是一个历史问题，而不是统计学问题。例如，历史学家可能认为没有理由相信他的资料受到任何经常的周期性因素的影响，并认为他的时间数列仅受到长期趋势或非经常性波动的影响。如果他确信如此，那么他就不应使用上述方法去排除并不存在的周期性因素，并且他也可以忽略以下几段的论述。

然而，如果我们稍有理由相信时间数列中存在着周期性成分，那么我们只有将它分离出来才能对此加以研究，也只有在我们从时期数列中排除了趋势和周期性成分之后才能对剩下的非经常性波动进行研究。从数列中排除周期性成分的最通用的方法被称为移动平均数方法。移动平均数方法的步骤和结果可从表6.8和图6.9中看出，这里显示的是一个有着绝对经常的周期性波动的假设资料集，其中每4年一个高峰，每4年一个低潮。若如表6.8所示，我们取最初4个数值的算术平均数，再取第2—5个，然后取第3—6个数值的平均数等等，我们就得到一个绝对经常性的和线性的数列，也即其中没有波动。因此，通过移动平均数我们就排除了周期性成分的影响。

然而，这种方法也常碰到一些通常不为历史学家充分认识到的困难。对于表6.8中的资料，我们通过以4年为一个阶段的移动平均数排除了周期性成分。我们这样做而不是例如以3年为一个阶段来取平均数，是因为假设的数列具有一个绝对经常性的4年周期(从高峰到高峰或从低潮到低潮的距离)。因此，当我们知道了这一周期的周期性，这种方法是行之有效的；我们于是可以选择移动平均数来适应这一周期性。

表6.8 计算移动平均数的方法：假设资料

时期	资料值	4年的总计	4年的平均数	5年的总计	5年的平均数
0	6				
1	5	20	5		
2	4	20	5	26	5.2
3	5	20	5	25	5.0
4	6	20	5	24	4.8
5	5	20	5	25	5.0
6	4	20	5	26	5.2
7	5	20	5	25	5.0
8	6	20	5	24	4.8
9	5	20	5	25	5.0
10	4	20	5	26	5.2
11	5	20	5	25	5.0
12	6	20	5	24	4.8
13	5	20	5	25	5.0
14	4	20	5	26	5.2
15	5				
16	6				

注意：N年的总计和移动平均数值，按惯例被置于所计算时期的中点的相对位置上。

然而，在大多数历史实例中，准确无误地决定周期性十分困难。例如，19世纪的商业周期的长度从5年到10年不等。我

们似乎只要确定某种平均的周期长度,比如说9年，并以此为基础进行移动平均数的计算。但不幸的是，移动平均数的方法的奏效在很大程度上取决于选择什么样的周期时间，选择错误可能导致极端使人误解的结果。这一点从表6.8中可以看出。若由于某种原因,我们取资料的5年移动平均数，那么我们就不可能得到一个平坦的数列。而且,5年移动平均数数列中的高峰将与原始数列中的低潮相对应，而前者的低潮则与后者的高峰相对应；以移动平均数为基础的数列将会完全错误地反映原始数列。

在图 6.9 的例子里，很明显已给人造成了一个错误的印象。可是在历史实例中,这一点可能表现得明显得多,而想要运用移动平均数方法的历史学家要时常警惕这种可能性，即在排除周期性成分的过程中,他正在歪曲余下的时间数列。

(如前所示,移动平均数的另一个困难在于它可将经常性波动引进一个实际并不存在经常性波动的数例。需要进一步考查这种可能性对历史时间数列的影响;与此同时,历史学家在将移动平均数方法运用于长期的时间数列时则更要特别小心谨慎。)

如果历史学家确信数列中存在着经常性周期，并能清楚地认定这一周期的周期性，那么应用移动平均数方法非常适宜。根据历史假设，1820—1850 年存在着一个经济活动的 5 年周期,它影响着英国本土的出口。据此,表6.9(第3、4、5列)显示了一个应用于表6.1中资料的5年移动平均数。表中第 5 列给出数列的周期性成分，从离差数列(第3列)减去第5列得出一个残差数列(第 6 列)。

在讨论这一残差数列之前，我们应指出移动平均数方法

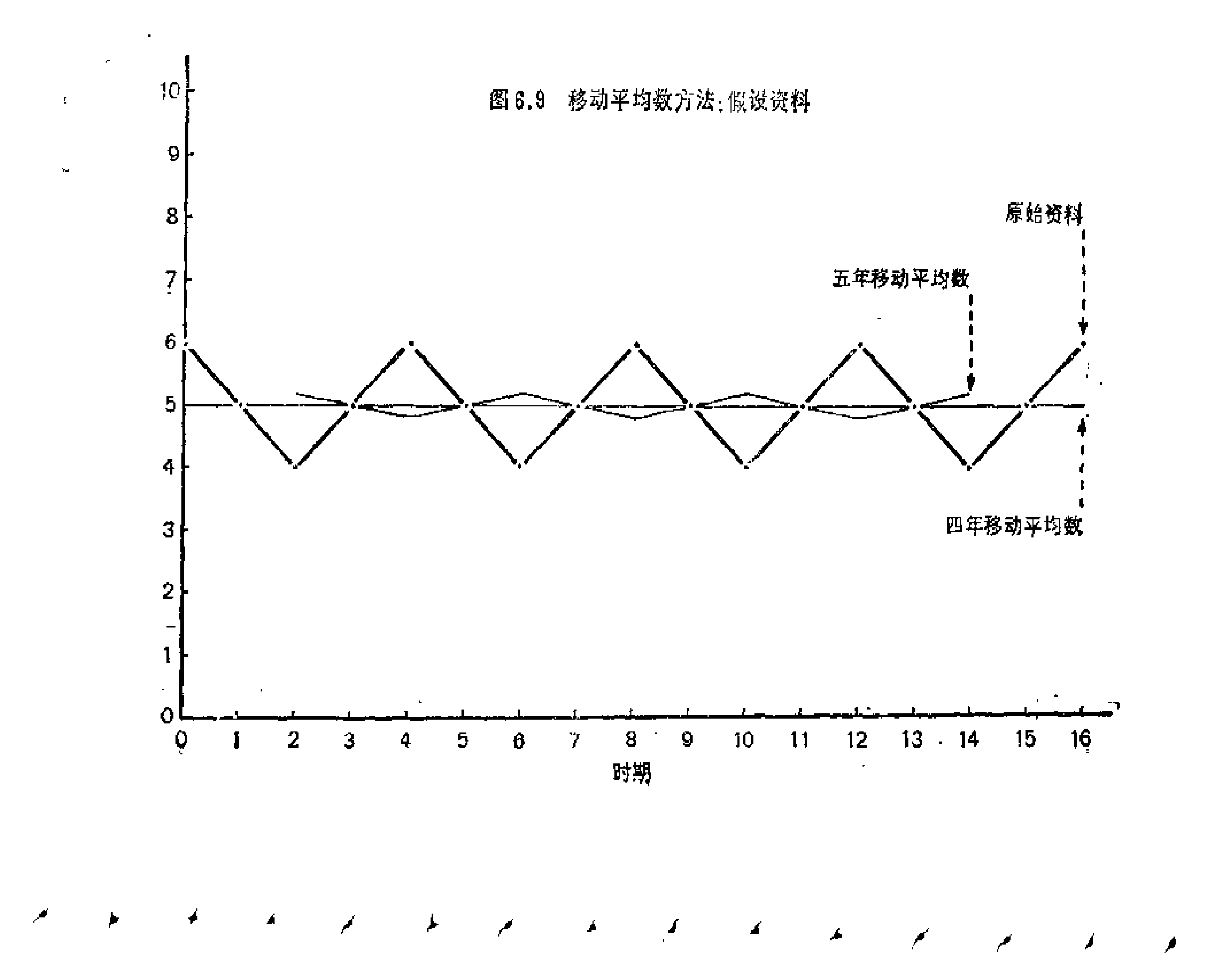


图6.9　移动平均数方法:假设资料

表6.9　将表6.1中的时间数列分为3种影响的方法

年份	I 资料值 (I)	II 趋势值 (II)	III 偏差 (I-II)	IV 5项偏 差之和	V III的移动平均 数:周期性波动	VI 非周期 性波动
1820	36.4	30.8	+5.6			
1821	36.7	31.8	+4.9			
1822	37.0	32.9	+4.1	+19.6	+3.9	0.2
1823	35.4	33.9	+1.5	+17.0	+3.4	−1.9
1824	38.4	34.9	+3.5	+6.7	+1.4	+2.1
1825	38.9	35.9	+3.0	+1.8	+0.4	+2.6
1826	31.5	36.9	−5.4	−1.9	−0.4	−5.0
1827	37.2	38.0	−0.8	−9.6	−1.9	+1.1
1828	36.8	39.0	−2.2	−15.3	−3.1	+0.9
1829	35.8	40.0	−4.2	−14.7	−2.9	−1.3
1830	38.3	41.0	−2.7	−20.5	−4.1	+1.4
1831	37.2	42.0	−4.8	−22.7	−4.5	−0.3
1832	36.5	43.1	−6.6	−22.0	−4.4	−2.2
1833	39.7	44.1	−4.4	−18.0	−3.6	−0.8
1834	41.6	45.1	−3.5	−7.0	−1.4	−2.1
1835	47.4	46.1	+1.3	−6.5	−1.3	+2.6
1836	53.3	47.1	+6.2	−1.2	−0.2	+6.4
1837	42.1	48.2	−6.1	+5.3	+1.1	−7.2
1838	50.1	49.2	+0.9	+4.2	+0.8	+0.1
1839	53.2	50.2	+3.0	−2.6	−0.5	+3.5
1840	51.4	51.2	+0.2	−2.4	−0.5	−0.7
1841	51.6	52.2	−0.6	−3.5	−0.7	+0.1
1842	47.4	53.3	−5.9	−3.2	−0.6	−5.3
1843	52.3	54.3	−2.0	+0.4	+0.1	−2.1
1844	58.6	55.3	+3.3	+1.5	+0.3	+3.0
1845	60.1	56.3	+3.8	+7.8	+1.6	+2.2
1846	57.8	57.3	+0.5	+3.2	+0.6	−0.1
1847	58.8	58.4	+0.4	+3.1	+0.6	−0.2
1848	52.8	59.4	−6.6	+9.3	+1.9	−8.5
1849	63.6	60.4	+3.2			
1850	71.4	61.4	+10.0			

注意：第2列(趋势)、第5列(周期性波动)、第6列(非周期性波动)相加为第1列(原始资料)。

还有一个缺点，即它在时间数列的初期和末期没有给出若干年的数值。当所用的是一个很长时期的移动平均数时，这点尤为严重。贝弗里奇为一些农业价格数列所用的31年移动平均数丧失了数列中前15年和最后15年的信息。这一缺点是否足以排除运用移动平均数方法，取决于所研究的时间数列的特征，并取决于这一数列初期值和末期值的重要性。

从数列中排除了周期性成分后，我们只剩下一个残差数列，如表 6.9 中的第 6 列所示。由于我们已将趋势和经常性波动从数列中排除，这就是数列中的非经常性波动。如果我们认为不可能识别出数列中的周期性波动，那么我们应将表6.9中的第3列视为由非经常性波动构成的。由于无法进一步再对数列进行简化处理，历史学家就必须运用其他技巧和历史知识来解释这些非经常性波动所以发生的原因。例如，1836年大幅度增长的波动和1837年下降的波动可以紧密地与美国经济的繁荣及突然的暴跌联系起来；1836—1837年英国对美国的出口削减了三分之二。

我们结束这一节对时间数列中波动的分析时，应当再一次强调，历史学家必须保证时间数列分析的假设与他自己所关怀的特定时间数列所作的历史假设相配合。这一必要性在运用移动平均数方法时表现得最为明显，为了这个原因，移动平均数的一个共通用途——获得对有关数列中趋势的认识——的有效性非常值得怀疑，除非可以假定一个周期性已知的周期存在于这一数列之中。然而，如果时间数列分析的假设得以成立，那么这些方法就非常有用，它使历史学家能够研究时间数列中的不同成分，并分别对长期和短期、经常性和非经常性的运动作出清晰的解释。

6.5 比率和指数的运用

在分析一个时间数列时，将数列中的每一个值用一个年度的值（通常是第一年的值）的比率来表示常是有用的。这样做，我们可以很容易知道数列中在发生些什么比例变化。例如在查考英国出口资料时，了解1830年，1840年和1850年的出口值以什么比例超过1820年的出口值是有用的，从而对出口值的增长得到一个印象。我们只要用1830年，1840年和1850年的每个值除以1820年的值便可以计算出这些比例，其结果如表6.10所示。

表6.10 作为1820年出口值比例的英国本土出口

年份	原始值（百万英镑）	原始值除以1820年值
1820	36.4	1.0000
1830	38.3	1.0522
1840	51.4	1.4121
1850	71.4	1.9615

如果我们愿意，还可以用1820年的比例来表示数列中的每一个值，从而建立起一个新的时间数列，其值始于1820年的1.0000而终于1850年的1.9615。我们还可以，而且这样做将更为正规，不以比例而是以1820年值的百分比来表示这些值。这样1820年的值为100，1830年为105.2，1840年为141.2，1850年为196.2。

我们称这一过程为"以1820年＝100为基年，将数列转换成比率形式"。以这种形式表示数列有许多优点，主要是用比率比我们必须用心算把51.4作为36.4的一个比例较为容易评价

比例变化。当时间数列的单位难于处理时，这一点就特别有价值；比如，工资标准常以先令或便士为单位表示，当它们以比率数列的形式表示时，评价它们的变化就容易得多。比率的另一个优点在于它便于把两个数列进行比较。例如，如果我们想要对1820—1840年的出口增长和同一时期的进口增长进行比较，那么比较出口从100增长到141.2而进口从100增长到168.3，较之比较出口从36.4（百万）英镑增长到51.4（百万）英镑而进口从54.2（百万）英镑增长到91.2（百万）英镑更为简易。

表6.11 不同基年的英国本土出口比率数列

年份	原始值（百万英镑）	1820＝100	1830＝100	1840＝100	1850＝100
1820	36.4	100	95.04	70.82	50.98
1830	38.3	105.22	100	74.51	53.64
1840	51.4	141.21	134.20	100	71.99
1850	71.4	196.15	186.42	138.91	100

很明显比率的值取决于基年的值。从相同的原始数列中导出的各种比率数列随所选择的基年而不同。表6.11显示选择不同的基年对表6.10出口值数列的影响。

从表6.11可以明显看出，不仅比率值随基年选择的变化而变化，而且，作为这种方法的自然结果——值与值的区间也随基年选择的变化而变化。因而说出口值的区间为2.66（基年1850年＝100）和该区间为5.22，4.96和3.69是同样地正确。基年的值越高，在比率数列中较低值年份之间的区间也就越小。

当我们选择基年并以此来计算数列值时，能够意识到这一点极为重要。在具有上升趋势的英国本土出口这类数列里，

选择早期年份为基年会给人一种数列迅速增长的印象（如从100到196.15），而选择较晚期年份似乎又会使这一增长减少（如从 50.98 到100）。但事实上，在比例变化方面两者并无差别，只是造成了一个不同的印象——至少对漫不经心的读者来说是如此。同样，在一个波动很大的数列里，随着所选择的基年是否为趋势中一个相当向上波动或向下波动的年份，比率数列会给人以不同的印象。

因而，基年的选择对于比率数列的应用至关重要。很难说对选择问题有一个准确答案，虽则一般说来我们应选择一个接近趋势线的年份值。选择一个靠近数列中间的年份也是明智的，但不幸的是这两个要求可能会发生冲突。此外，由于需要运用比率去比较两个或更多的数列也常常使这个问题复杂起来。遇到这种情况，我们只能为所有数列选择一个使任何一个数列的失真度降至最小的基年。这是一项非常棘手的工作，并且对此无法定出规则。

比率的这些欠缺必须与应用比率所能造成的清晰的真实好处相对比。这些欠缺可以用连同指数一起来规定原始值予以减少，使读者可以知道以指数形式来呈现，在多大程度上正改变他对资料的印象。

在一节里迄今我们已考虑了将一个或多个时间数列分别转换成比率形式的情况。比率的另一个用途是作出联系几个不同时间数列的综合指数，这或许更为重要。这类指数中最为人熟知的例子是零售物价指数，其他在历史学研究中用到的是如工资标准指数，生活费用指数，等等。在考虑诸如工业革命时期中生活水准是否已改善或者下降这类问题时，这些指数极为重要。为回答这类问题，就需将生活费用与收入或工

资相比较，而要这样做就需作生活费用指数和工资指数。

对于作这类指数的方法，及其在历史学和统计学方面所涉及的困难的一个例子，我们将考虑 1890—1900 年生活费用指数的制订工作。"生活费用"，这个名词包括用于食品、房租、衣服、燃料，及杂物等方面的支出，所以我们必须考虑到每一项费用的变化。另外，其中每一项本身都由若干不同的费用组成。我们还必须考虑到这样的事实，即面包价格可能以不同于肉、鱼和其它食品费用的方式变化。然而为了清晰起见，让我们假定，我们已经编制了生活费用的主要成分的指数，如表 6.12 所示。

表 6.12 1890—1900年生活费用指数成分（基年1900＝100）

年份（权数）	食品（60）	房租（16）	衣着（12）	燃料（8）	杂项（4）	综合指数
1890	101	93	102	80	89	97.68
1891	103	94	102	78	85	98.72
1892	104	95	101	78	81	99.20
1893	99	96	100	85	81	96.80
1894	95	96	99	73	75	93.08
1895	92	97	98	71	75	91.16
1896	92	98	99	72	75	91.52
1897	95	98	98	73	75	93.28
1898	99	99	97	73	74	95.68
1899	95	99	96	79	76	93.72
1900	100	100	100	100	100	100.00

资料来源：A. L. 鲍利：«1860 年以来英国的工资和收入» (A. L. Bowley, Wages and Income in the United Kingdom Since 1860)，剑桥：剑桥大学出版社，1937 年，第 120—121 页。

我们的任务是将这些不同的生活费用指数合并成一个指数。一种方法是简单地取每一年不同指数的算术平均数。如

在 1890 年，我们得到的是

$$\frac{100+93+102+80+89}{5}=\frac{465}{5}=93。$$

这一做法的困难在于我们正试图编制一个将使我们能得到有关 1890—1900 年实际人民的生活费用变化的印象。然而人们并非都是将他们家庭收入平均地花费到 5 种生活费用中的每一项上。大多数人在食品上的开销超过在其他需求上的开销，而且我们当然没有理由认为在衣着上的支出应与在燃料上的支出一致。所以，为使我们能够将生活费用指数应用到1890—1900年生活水准变化的问题中，我们必须考虑到家庭预算中不同支出的不同重要性。因此只取这 5 项的算术平均数是不能令人满意的；我们反之需要更多地考虑食品价格的变化而不是燃料价格变化的变化，因为食品价格的变化将具有更大的影响。

我们通过给 5 项中的每一项指定不同的"权数"来达到这一目的。表 6.12 显示了这些权数。基本上，我们认为这一时期的中等家庭在食品上的支出是在衣着上支出的 5 倍，因此食品价格上的变化应被认为比衣服价格上的变化重要 5 倍。因此，为计算每一年的综合指数，我们将每一项的指数乘以它的权数，再除以加权总数（本例中为60＋16＋12＋8＋4＝100，但权数并不需要合计为100）就得到综合指数，如表 6.12 最后一列所示。例如，对于 1890 年，我们有(100×60)＋(93×16)＋(102×12)＋(80×8)＋(89×4)＝9768，除以100(权数的总数)后，得到97.68，这就是综合加权生活费用指数。

因而加权指数的计算不过是一种比较简单的算术运用。而建立这类综合指数的困难不在于其统计步骤，而在于对证

变化的历史学家必须能够发现“实际”工资，即依据工人必须购买商品的价格的变化调整过后的工资；换句话说，必须是按照它们的购买力来表示的工资。

在上述2个例子中，一组“货币”值，无论是价格还是工资的，都必须经过价格指数的缩减之后才产生一组“实际”值。我们可以以表6.12计算的生活费用指数为例了解一下缩减一个工资数列的情况。表6.13中第1列就是工资数列；它通过取不同职业的众多工资数列的加权工资平均数得出，其做法与生活费用综合指数的作出大致相同。此指数最初以1914年

表6.13 1890—1900年实际工资指数的建立

年份	货币工资(1914＝100)	货币工资(1900＝100)	生活费用(1900＝100)	实际工资(1900＝100)
1890	83	88.3	97.7	90.4
1891	83	88.3	98.7	89.5
1892	83	88.3	99.2	89.0
1893	83	88.3	96.8	91.2
1894	83	88.3	93.1	94.8
1895	83	88.3	91.2	96.8
1896	83	88.3	91.5	96.5
1897	84	89.4	93.3	95.8
1898	87	92.6	95.7	96.8
1899	89	94.7	93.7	101.1
1900	94	100.0	100.0	100.0

资料来源：货币工资指数来自 E. C.拉姆斯博登，转引自 B. R.米切耳和 P.迪恩：《英国历史统计摘录》，第346页，生活费用指数来自表6.12。

的100为基数进行计算，再以1900年为基年＝100重新计算，并在表中第2列显示出来；只要把每一值除以94（1900年的值），再乘以100就可做到这一点。表6.13中第3列显示综合的生活费用指数。我们将生活费用指数的每一值去除以货

币工资指数的相应值，再将结果乘以 100 就得到表中第 4 列的最终数字。

从表 6.13 可以明显看出，我们刚才算出的“实际”工资指数与货币工资指数相差很大。货币工资指数从 1890—1896 年是稳定的，接着开始上升至 1900 年，而实际工资指数从 1890—1892年是下降的，以后开始上升至 1895—1896 年，而后经过轻微的下降后再上升至 1900 年，研究这一时期工会史这类问题的历史学家必须认识到这个差异；仅知道货币工资率大概不能作为这一时期劳工史的研究的良好指导。

这一节里，我们只可能讨论了若干种最简单的建立指数的方法，以及其最普通的用法。正如这一节已介绍过的那样，这些建立指数的方法的差别不在于其基本的逻辑，而在于指定权数，选择基年，以及类似的问题。因此，面临涉及指数的困难问题的历史学家应根据这一节里已介绍的逻辑概念查阅本书开列的有关参考书之一。

7 变量之间的关系

在前几章里我们已经讨论了历史学家所使用的大部分计量方法；直到最近才有少数有关历史的著作和文章应用了比我们叙述过的诸如频数分布、集中趋势或离中趋势的测度方法，以及时间数列分析更为复杂的统计方法。人们应用这些统计方法已写出了大量重要的历史著作。可是一个打算运用计量方法的历史学家不应当仅停留在这一阶段上，而应进一步运用其他可以帮助他分析历史材料的、更为复杂的方法。在本书中不可能对所有这些方法加以讨论，因此在本章里我们将集中于探索在撰写历史中的主要问题的技术——即两组历史事件之关系的问题的方法。这些方法将被列入“相关与回归技术”的总标题之下加以讨论。

历史学家所讨论的很多问题可以被概括成是否存在着一种“关系”的问题。例如，我们想知道在我们对1907年航运业的研究中船员人数与动力类型之间，或1688年收入与社会地位之间，或1086年牧猪数量与草地面积之间，或下议院的一次投票结果与另一次投票结果之间，或19世纪英国出口与进口之间是否存在着一种关系。我们问是否存在着一种关系

的目的不过是想要了解两个或更多事件彼此之间是否完全不相干，或者是否它们之间存在着某种联系，不论微细到什么程度。判定了是否存在着一种关系后，我们就可以进一步充分了解这一关系的强度和它的形式。例如，我们问这种关系是否强烈得当A发生时B必然随之发生，或者较弱到在A发生时的大多数(但非一切)情况下B也随之发生。我们问这种关系是否具有A增加B也增加这种形式，或者是否关系正相反，即当A增加时B却减少了。

现有的大量统计方法可以帮助我们回答这类问题。然而重要的是认识，只有当我们运用了自己的历史知识提出有意义的历史问题时，这些方法才能帮助我们。例如，完全有可能用这些统计方法去检验下院投票与月相之间是否存在着一种关系，而由于巧合，两者之间很可能存在着某种统计上的联系。然而，这种联系没有任何历史上的意义，对历史学家来说只是一个没有价值的结果。问是否存在着这样一种联系本身就是个愚蠢的问题，而当然我们也只能得到一个愚蠢的结果。换句话说，在我们应用相关和回归法之前，我们必须能够明确我们怎样认为两个变量之间可能存在着联系，然后再看统计论据是否支持我们的理论；我们必须能够用历史学的和统计学的语言来描述这种可能的关系。

实质上，对于两个或更多的历史事件之间的关系我们可试图回答3个问题。它们是：

1. 是否有关系？
2. 关系的强度如何？
3. 关系为何种形式？

下面让我们来讨论能够帮助我们回答上述问题的统计方

法。

7.1 是否有关系？

让我们设想，作为一个历史研究的结果，我们认为一系列事件与另一系列事件相关联。换句话说，我们有这样一种形式的假设："我认为变量 1 与变量 2 相关联"。在某些情况下，这种假设可能并不重要而且毋庸证明；"在都铎王朝时代的英格兰叛国者的斩首和他们的死亡有关"的陈述就属于这类假设。大多数涉及关系的令人感兴趣的假设都不会是这种形式；因为它们不能根据人没有脑袋就无法生存的生理学法则来验证，而需要更复杂的证据。

通过将变量 1 与变量 2 相关联这一假设与变量 1 与变量 2 毫无关联的另一假设相对照，我们可以很容易处理怎样证明"变量 1 与变量 2 相关联"这一假设的真实性问题。这另一个假设与"变量 1 与变量 2 不相干"的假设是相等的，根据这个假设，我们的意思是说我们在变量 1 的一个值与变量 2 的一个值之间不期望找到关系（除了两者同属某一事例的微细关系之外）。另外一个相等的假设为"对同一事例，变量 1 的一个值在预测变量 2 的值时根本不能给我们帮助"。例如，我们可以用假设"船员人员与船的规模相关联"去对比下面另一假设——"船员人员与船的规模无关"，"船员人员与船的规模不相干"，以及"知道船员人数根本不能帮助我们预测船的规模"。

用这些替代的假设来重述我们的最初假设，其价值在于我们可以在调查两个变量之间的关系时进一步问："如果两个

变量确实彼此不相干，资料集会呈现出什么样子?”实际上，我们根据变量之间彼此不相干的假设建立起了另一个资料集；并用以与我们以两个变量是有关的首先假设为根据的实际资料集相对比。如果实际资料集与假设资料集看来很相似，那么我们或许就会得出如下结论：完全可以设想两个变量毫无关联。如果两个资料集相差悬殊，那么我们还是设想两个变量之间大概存在着某种关系为可靠。我们仍不知道这种关系为何种形式，只知道资料不支持没有关联的假设。以我们的商船研究为例，根据变量之间彼此不相干的假设我们建立起另一个资料集，并用它与实际资料集相对比；如果我们发现两者很不相同，那么我们便可以认为船员人数与船的规模之间存在着某种关系。此外，部分通过进一步的统计工作，部分根据历史知识，我们可以继续考查这种关系可能是由什么所引起的。

为了能够进行这一真正的资料集和假设的资料集的对比过程，我们需要做两件事。我们需要建立另一假设资料集，而且我们还需要判断这一假设资料集是否真正地不同于实际资料集。我们将讨论这样做的 2 种方法。第一种方法为计算列联系数 C，不论是定名、定序和区间类型的资料都可适用；第二种方法为计算相关系数 R，它只适用于区间类型的资料。我们不打算讨论其他几种适用于定序资料的方法；因为定序资料在历史研究中较为罕见，因此这些方法不大会常用。

然而应该指出，有时当资料貌似区间类型但又不能完全肯定时，应用适用于定序资料的方法是明智的。第一章中引用的格列高里·金有关收入和社会地位的资料就是一例。在这种情况下，定序方法可以作为一种安全措施来应用，其结果可

以同那些从适用于区间资料方法所得的结果相比较。有关对应用定序资料进行检验的信息可以很方便地在一本非常有价值的书中找到，它就是S.西格尔的《行为科学的非参数统计学》。

我们将首先考虑列联系数C的计算和解释；顾名思义，它最常用于决定已制成列联表形式的变量之间是否存在着某种关系。我们可以通过用一个取自英国政治史的简单例子来最清楚地说明其用途。在1841年选出的国会中，国会议员对各自政党（自由党和保守党）的忠诚相当强烈，但在某些问题上，其他的忠诚超越了政党的束缚。例如，在1845—1846年有关废除《谷物法》的问题上，很多保守党议员投票反对保守党领袖及总理罗伯特·皮尔。因此，去发现在国会对其他问题的投票是否遵循着政党路线是饶有趣味的。例如，我们可以调查1844年对有关棉纺厂童工每日工作时间是否应被限制在10小时以内问题的最后辩论和表决中，政党束缚是否决定着投票行为。在那次表决中，94名自由党议员和100名保守党议员投票赞成限制，而56名自由党议员和135名保守党议员投票反对。这次投票的情况在表7.1中以左边的列联表形式列出。

我们对在这个问题上议员投票是否与政党束缚有关感兴趣。因此我们的最初假设是政党束缚与投票相关联，而另一假设为政党约束与投票互不相干。为了在这两个假设之间进行抉择，我们需要以政党与投票无关的设想为根据而建立另一个投票型式；此后我们便可将实际投票型式与假设投票型式进行对比。

例如，如果自由党议员中投票赞成此议案的比例远远大

表7.1 1844年《10小时议案》的实际和假设投票

	观察投票数			期望投票数		
	赞成	反对	总计	赞成	反对	总计
自由党	94	56	150	75.6	74.4	150.0
保守党	100	135	235	118.4	116.6	235.0
总　计	194	191	385	194.0	191.0	385.0

资料来源：W. O. 艾德洛特："19世纪40年代英国下议院的投票型式"(W. O. Aydelotte, 'Voting patterns in the British House of Commons in 1840s')，载《社会和历史的比较研究》(Comparative Studies in Society and History)第5卷(1963)，第134—136页，表3。

于所有议员投票赞成它的比例，那么我们自然会猜想对自由党的忠诚很可能会产生赞成此议案的投票。对比之下，如果在这个问题上政党束缚根本没有影响投票行为，那么我们会期望发现，投票赞成此议案的自由党的比例将与所有投票赞成的议员的比例大致相同。这个根据常识的对比，提示我们该怎样建立起自己的假设投票型式，设想政党束缚与投票行为之间没有任何联系。

在这次表决中总共有385人参加了表决，其中194人即50.4%投票赞成，而191人即49.6%投票反对。在投票的385人中，自由党议员为150名，其余的235名为保守党议员。先来看自由党议员，如果自由党束缚与此次议案的投票毫无关联，那么我们期望大约应会有50.4%的自由党议员投票赞成此议案，而49.6%的自由党议员反对。

算出150的50.4%为75.6，我们认为，根据两变量之间彼此不相干或没有联系的假设，应会有75.6名自由党议员投票赞成此议案而不是实际投票赞成的94名。75.6这个数字被称为投票赞成此议案的自由党人的"期望值"（在彼此不相

干的假设下所期望的)；94 这个数字被称为“观察值”。

然后我们可以像表 7.1 那样计算其他可能的政党束缚与投票行为的组合的期望值，并可将其结果整理成列联表的形式，对比“期望的”和“观察的”投票型式。注意在表的左右两边我们计算并给出了行与列，以及它们的总和。这有两个用处：其一，它使我们可以通过保证行和列的总数与在原始资料中的相一致，检查期望值的计算；其二，它为我们提供了一个计算期望值的简便方法。这是因为表中每一单元的期望值都可以通过它所在的行的总数乘以它所在的列的总数，再除以总和的结果而计算出来。例如，对投票赞成此议案的自由党议员来说，其期望值得出为

$$\frac{150 \times 194}{385} = 75.6$$

这与我们以前所获的结果一样。

现在我们已经完成了任务的第一步，即根据投票行为与政党束缚之间没有关系的假设建立另一个假设的资料集。第二步我们应比较观察和期望的投票数，以决定哪一个假设(即变量之间有关系或变量之间没关系)最令人满意。我们通过用表中每一单元中的观察投票数减去表中相对应单元中的期望投票数，将其平方以去掉负号，再除以此单元中的期望值并以一种相对的形式表达其结果来做到这一点。然后将所有结果求和，得到一个被称为 χ^2 的量，读为“卡方”，记为“χ 方”。对于表 7.1 其计算结果为

$$\chi^2 = \frac{(94-75.6)^2}{75.6} + \frac{(56-74.4)^2}{74.4} + \frac{(100-118.4)^2}{118.4} + \frac{(135-116.6)^2}{116.6} = 14.8$$

因此 χ^2 公式的一般形式为

$$\chi^2=\sum_{i=1}^{R}\sum_{j=1}^{C}\frac{(O_{ij}-E_{ij})^2}{E_{ij}}$$

这里 R 是行数，C 是列数，i 为行下标，j 为列下标，O_{ij} 和 E_{ij} 分别为每一单元的观察值和期望值。

在像表 7.1 这样只有两行两列的列联表的特定情况中，上述的一般公式会产生一个夸大了的卡方。因此我们必须使用另一个更简便和更准确的公式。如果我们像表 7.2 那样标记一个 2×2 列联表的单元，卡方可用以下公式得出

$$\chi^2=\frac{N\left(|AD-BC|-\frac{N}{2}\right)^2}{(A+B)(C+D)(B+D)(A+C)}$$

(在这个公式里，$|AD-BC|$ 表示 $AD-BC$ 的"绝对值"；即我们不顾符号并把整个项视为正数，即使 BC 大于 AD)。

对于表 7.1 中的资料，用此公式计算的卡方为

$$\chi^2=\frac{385(|(94\times135)-(56\times100)|-385/2)^2}{(150\times235\times194\times191)}$$
$$=14.02$$

这另一公式必须用于 N 大于 40 的 2×2 列联表。如果这个条件不能得到满足，应查阅西格尔[①] 书中的另一种方法。

由于我们有一个 $N=385$ 的 2×2 列联表，在进行下一步计算时，我们将应用一个卡方等于 14.02 的值计算列联系数 C 通过一个简单的公式将列联系数 C 与卡方联系起来

① S. 西格尔的《行为科学的非参数统计》(Nonparametric Statistics, for the Behavioural Sciences),纽约:麦可喜图书公司，1956 年，第 110 页。

表7.2 标记2×2列联表的单元

A	B	$A+B$
C	D	$C+D$
$A+C$	$B+D$	N

$$C=\sqrt{\frac{\chi^2}{N+\chi^2}}$$

因此,对于我们的资料

$$C=\sqrt{\frac{14.02}{385+14.02}}=0.1875$$

现在我们必须以我们想要发现政党束缚与对《10小时议案》的投票行为是否存在着某种关系这一点出发来解释这一结果。如果我们回到计算卡方的第一个公式,可以看出,若实际资料与假设资料完全相同,以致观察值等于期望值,χ^2的值将为零。如果我们来看计算C的公式,若卡方为零,那么C也将为零。

实际上我们已经得到一个非零值的C,表明实际投票型式与假设投票型式彼此不同。因而投票型式不支持政党约束与投票行为无关这另一假设,我们也可以断定在有关《10小时议案》的表决中投票行为与政党束缚之间存在着某种联系。应该指出,迄今我们还没有考虑过这一联系强烈到什么程度和它的形式如何,也没有考虑过可能与投票行为有联系的其他因素;在本章的后面我们将回到这些问题上来,现在我们只关心是否存在着某种联系。

因而列联系数提供了一种使我们可以用一个建立在变量之间没有联系的假设之上的资料集与实际资料集相对比的方

法。它是一个可以广泛用于不同类型资料的测度方法。然而，当资料为区间类型时，我们可以换用另一种方法——相关系数。这一测度方法应用了区间资料给我们的额外信息，而且它在较高级的统计工作中亦有很大用处；由于这两个原因，当我们的资料为区间类型时，宁愿选用 R 而不是 C。

我们将通过第四章里的商船例子来考虑 R（其全称为"皮尔逊积矩相关系数"）的计算。在商船的变量特征中，我们列举了船的吨位和船员人数。研究这一时期（1907年左右）航运史的学者会知道其间存在着一个商船平均规模增加的趋势。这一趋势有可能增加了对船员的需求，即较大的船需要更多的人来操纵；另一方面，也有可能这一增加的规模大部分为载货舱位，而对船员需求的增加并不像对岸基货物搬运工需求的增加那么大。因此，考查表 4.1 的资料中船的吨位与船员人数之间是否存在着一种联系是令人感兴趣的。

最初假设为存在着这样一种关系。另一种假设为吨位与船员人数互不相干。首先让我们考虑，对于每一个假设我们会期望资料呈现出什么样子。如果船员人数与吨位相关联，那么我们会期望对有关一艘特定商船的变量之一的了解，会使我们得到这艘船另一变量的大概信息，不论它是多么粗略。例如，我们大致期望大船拥有大量船员，小船拥有小量船员。对两艘船进行比较，一艘船的吨位为另一艘船的两倍，那么我们甚至可以期望它拥有的船员也比另一艘船大致多两倍。

与此对照，如果船员人数与吨位完全不相干，我们则不会期望一个变量的高值会与另一个变量的高值相联系。确实，对于每一个吨位值，我们会期望出现船员人数的分散值，有的较高，有的较低，有的居中。了解吨位甚至不会给我们以任何

方式的帮助去猜测船员人员的多少。

当我们在区分这两个假设时，我们已描述了一种两个变量相关联的可能情况，在这种情况中，如果吨位高时，船员人数也高；对于另一种假设，则如果吨位高，船员人数可能高，也可能低。但立即生起的问题是："高或低意味着什么"、"我们怎样衡量这些相对的概念"？我们记得，一种判断一个变量的特定值是高还是低的方法，即是发现它是否高于或低于此变量的平均数，以及偏离平均数多远。因此，我们可以换一种说法来说明上述两项假设：第一，如果两个变量相互关联，我们会期望高于吨位平均数的吨位会与高于船员人数平均数的船员人数相联系。第二，如果两个变量之间彼此不相干，那么高于吨位平均数的吨位既很可能随伴低于船员人数平均数的船员人数也很可能随伴高于它的船员人数。根据变量之间彼此不相干的假设，部分资料集可以看来像表 7.3 所表示的那种情形。

表7.3　根据变量之间彼此不相干的假设，假定的商船资料

商船	A：与吨位平均数的关系	B：与船员数平均数的关系	$A\times B$
1	高于平均数（+）	低于平均数（−）	一个负数
2	等于平均数（0）	等于平均数（0）	零
3	高于平均数（+）	高于平均数（+）	一个正数
4	低于平均数（−）	低于平均数（−）	一个正数
5	低于平均数（−）	高于平均数（+）	一个负数

对于每一艘船，若我们用它与吨位平均数的偏差 A 乘以它与船员人数平均数的离差 B，我们将有时得到一个正量，有时一个负量，这取决于表7.3中所表示的哪种情况适合于一艘特定的商船。在长期中，这些正的和负的量将趋于相互抵销，

以致正的和负的量之和，即种种平均数偏差的积将接近于零。

另一方面，如果两个变量之间相互关联，我们或许会得到一个如表 7.4 表示的情形。

同样，在长期中，一个如表 7.4 所表示的情形将所有偏差求和以后会产生一个很大的正数值。

表 7.4　根据变量之间有关系的假设，假定的商船资料

商船	A：与吨位平均数的关系	B：与船员数平均数的关系	A×B
1	高于平均数（+）	高于平均数（+）	一个正数
2	高于平均数（+）	高于平均数（+）	一个正数
3	等于平均数（0）	等于平均数（0）	零
4	低于平均数（−）	低于平均数（−）	一个正数
5	高于平均数（+）	高于平均数（+）	一个正数

这提示决定两个变量之间是否有联系的一种方法，就是检验每一个案中每一变量与它们各自的平均数的偏差之积，并将所有个案积求和的结果。如果此结果接近于零，则很可能两变量之间没有联系，但是如果结果远离零，则有理由认为存在着一种关系。

因此，为了发现两个变量是否相关联，我们可以从看每一个案中的每一变量与平均数的偏差开始。我们将用一个公式进行计算

$$\Sigma(X-\overline{X})(Y-\overline{Y})$$

若计算结果为零，我们大可假定两变量之间没有联系。然而，若计算结果不为零，将会出现个案越多结果很可能越大的困难；如果变量用数百万而不是数百来表示，其结果也会较大，虽然比例偏差可能无异。为便于不同资料集之间的比较，我们可以把结果先除以个案数，再除以每一变量的两种标准差

之积。这样就消除了具有不同个案数的影响，以及某些围绕平均数有一较大散布值的资料的影响。

下面是计算积矩相关系数R的公式：

$$\frac{\Sigma(X-\overline{X})(Y-\overline{Y})}{Ns_x s_y}$$

这里 s_x 为一变量的标准差，s_y 为另一变量的标准差。如果两变量之间没有关系将得到一个零值；如两个变量之间有关系将得到一个大于或小于零的值。

如果记得X的标准差是得自

$$s=\sqrt{\left(\frac{\Sigma(X-\overline{X})^2}{N}\right)}$$

我们可以重写求R的公式为

$$R=\frac{\Sigma(X-\overline{X})(Y-\overline{Y})}{N\left[\sqrt{\left(\frac{\Sigma(X-\overline{X})^2}{N}\right)}\right]\left[\sqrt{\left(\frac{\Sigma(Y-\overline{Y})^2}{N}\right)}\right]}$$

这个公式还可以写成更简便的形式，不再需要计算平均数的偏差，如

$$R=\frac{N\Sigma XY-\Sigma X\Sigma Y}{\sqrt{([N\Sigma X^2-(\Sigma X)^2][N\Sigma Y^2-(\Sigma Y)^2])}}$$

而表 7.5 显示从表 4.1 的商船资料中计算R的过程。

如表 7.5 所示，商船资料的R为非零值，因此我们可以断定实际商船资料不同于根据两个变量之间彼此不相干的假设所建立起来的假定资料。在计算R的过程中，事实上我们并没有像计算列联系数C那样建立起另一个资料集，但原理是一样的——用一个假定资料集与一个实际资料集相对比。对于这两种测度方法，实质上我们所问都是："我们现有的资料集是否与若干变量之间彼此不相干时我们会期望的资料集不同？"

表 7.5　根据表4.1中的资料计算相关系数 R

官方号码	船员人数 Y	吨位 X	Y^2	X^2	XY
1697	3	44	9	1936	132
2640	6	144	36	20736	864
35052	5	150	25	22500	750
62595	8	236	64	55696	1888
73742	16	739	256	546121	11824
86658	15	970	225	940900	14550
92929	23	2371	529	5621641	54533
93086	5	309	25	95481	1545
94546	13	679	169	461041	8827
95757	4	26	16	676	104
96414	19	1272	361	1617984	24168
99437	33	3246	1089	10536516	107118
99495	19	1904	361	3625216	36176
107004	10	357	100	127449	3570
109597	16	1080	256	1166400	17280
113406	22	1027	481	1054729	22594
113685	2	45	4	2025	90
113689	3	62	9	3844	186
114424	2	68	4	4624	136
114433	22	2507	484	6285049	55154
115143	2	138	4	19044	276
115149	18	502	324	252004	9036
115357	21	1501	441	2253001	31521
118852	24	2750	576	7562500	66000
123375	9	192	81	36864	1728

$\Sigma Y=320$ $\Sigma X=22319$ $\Sigma Y^2=5932$ $\Sigma X^2=42313\ 977$ $\Sigma XY=470050$

$$R=\frac{N\Sigma XY-\Sigma X\Sigma Y}{\sqrt{([N\Sigma X^2-(\Sigma X)^2][N\Sigma Y^2-(\Sigma Y)^2])}}$$

$$=\frac{25(470050)-(320)(22319)}{\sqrt{[(25(42313977)-22319^2)(25(5932)-320^2)]}}$$

$$=0.9093.$$

如果C或R为非零值，那么这个问题的答案将为“是”，而且我们否定变量之间彼此不相干的假设，而赞成变量之间有联系的假设。

7.2 关系的强度如何？

在上一节里我们讨论了如何确定两个变量之间是否有关系。正如我们指出的那样，它并没有详尽无遗地回答有关变量之间关系的所有问题。我们感兴趣的第二个问题是：“关系的强度如何？”

对于这一问题简单的，却过于简单的，答案即是C或R与零相差越远两变量之间的关系越强。尽管这是对的，但它并不是对问题的一个完整的答案，因为若两个变量绝对相关，人们还想知道C或R的值是多少；“绝对相关”意谓两个变量之间存在着某种固定的关系，如知道了某一特定个案中的一个变量值就能使人精确地知道此个案中另一个变量的值是多少。规定出了C或R可能达到的值域，即从彼此不相干到全部相关，人们就能判断出任何一个特定资料集中的关系的强度，还能以各自所展现出来的变量之间关系的强度来比较两个资料集。

如果我们考虑列联系数C，我们发现，正如我们知道的，当变量之间彼此不相干时会产生一个零值。不幸的是，C的最大值取决于列联表的规模。对于一个2×2的列联表，变量之间的绝对相关将使$C=0.707$；对于一个3×3的列联表，C的最大值为0.816。了解了这一点我们便可以说，在我们的例子里，政党约束与对《10小时议案》投票行为的联系，$C=$

0.1875，是相当微弱的，在2×2列联表C的允许值域中处于很低的位置。然而，我们不能用一个2×2的列联表的C去和一个3×3的列联表的C相比较，而且我们不知道行数与列数不等的列联表C的最大值是多少。这些是列联系数和基于X^2来研究变量之间关系程度的其它测度方法的严重缺陷。因此在本文中叙述列联系数并不是由于它是一个理想的测度方法，而是由于它也许是最常使用的测度方法，并由于在它的计算中的逻辑构成了许多其他检验方法的基础。想要应用这类检验方法的历史学家，因此极应考虑参考书目开列的著作中所叙述的其他检验方法。

然而，如果我们的资料为区间型，则较为幸运，因为相关系数R的允许值域是被清楚地限定出来的；若变量之间没有关系R将为零，若变量之间是绝对的和正的相关，R的值则为+1，若变量之间是绝对的和反的相关，则R的值为-1。正相关的含义为如果一个变量有一个高值，很可能另一个变量也会有一个高值；反相关的含义为如果一个变量有一个高值，另一变量却有一个低值，反之亦然。由于R的允许值域如此清晰，我们可以说从资料中得到的值越接近于+1或-1，变量之间的关系就越密切。并由于R的值域不取决于个案的数目，我们还可以直接用一个资料集的R值同另一个资料集的R值相比较。

我们有关商船吨位和船员人数的资料已得出一个R等于+0.9093的值。它表明在我们所用的资料中，这两个变量之间存在着一种相当强的正关系。如果这些资料对1907年的所有商船具有代表性，那么我们可以说对于1907年的所有英国商船来说，吨位与船员人数之间存在着一种很强的关系；不

过，这是下一章里所要考虑的一个单独的问题。这里，必须强调指出，一个资料集中一个特定R的存在，并不意味着相同的R，或者一个相同的关系，存在于所有同类的资料之中。

由于来自不同资料集的R值可以直接进行比较，我们通过计算其它年份吨位和船员人数之间的相关系数，可以继续对商船进行考查，以观察吨位与船员人数之间的关系是否随着时间推移而加强或减弱。然而，在做这项工作之前我们应该对R及R的不同值的确切解释十分清楚，在下一节里对此将进一步加以讨论。

7.3 关系的形式

在若干历史问题中，仅仅试图证实两个变量相关，并判断出这一关系的强度就足够了。只要这一关系相当强，而且历史学家相信这并不是一种巧合（我们在下章谈这一问题），他就可以运用自己的历史知识去解释这一关系的历史含义。然而，在很多事例中，注意力并不集中于关系的存在，这可以假定或是自明的，而倒要集中于关系的确切形式。

例如，一位研究经济史的学者可以正常地假定，一种特定商品的制造量与它的出售价格之间存在着某种关系。因此他将不特别感兴趣于证实这一关系的存在，而将想要查考价格怎样精确地随着出售的商品数量而变动。与此对比，一个研究19世纪政治史的学者会认为证实国会议员按照政党束缚而投票其本身就是对知识的一个重要贡献。从这些例子可以看出，确定关系的哪一个特征是感兴趣的，其主要因素之一就是人们对此关系在理论上和经验上的认识情况。然而，可

以认为所有研究，无论从什么理论或经验的基础出发，其目标都是尽可能多地去发现，而在这一节的其余部分我们将在这基础上继续讨论。

首先让我们问何为两个变量之间关系的形式。我们是意指两个变量发生关系的方式，通过回答下列问题我们将发现它："关系为正相关还是反相关?"，"变量X需变动多大才能在变量Y中产生变化?"，"能否以变量X的变化来解释变量Y的所有变化，或许还牵涉到其他因素?"

如果我们的资料为定名类型，已在列联表中分类并用列联系数或相似的测度方法分析过，那么上述问题或则不适用，或则不可能为统计分析的任何具体方法所解答。与相关系数不同，列联系数总是一个正值，因而其计算不能告诉我们两个变量之间的关系为正相关还是反相关。但事实上并不需要用一个试验来告诉我们这一点；如果想知道作为一个自由党议员是否与投票赞成《10小时议案》正相关，只需直接观察投票结果即可。然而，若问总的来讲政党束缚是否与对《10小时议案》的投票正相关或负相关却毫无意义；这两个变量既可以相关，也可以无关，而我们用已经讲过的统计检验方法去查明这点。

然而，统计分析可对确定两个定名尺度变量之间的关系的形式是重要的一个问题给予很大帮助。我们已用这一提问考虑过有关《10小时议案》的投票："政党束缚在决定对此法案的投票是否重要?"然而在肯定是重要时，我们并未考虑其他对议员起作用的因素是否也具有同等或甚至更为重要的意义。事实上，我们可以期望政党束缚将是议员需加考虑的一个因素，但他也会有一定的"思想"见解，它有时会使他反对自

己的政党。出于对保护英国农业的重要性的关心，1846 年很多保守党议员反对皮尔，就是一例。假如作为个别或集团的议员除了对他们的政党忠诚以外还有共同的“思想”，那么就有可能通过研究议员们对一系列问题的投票而认明这些思想。帮助解决这类历史问题的方法被称为古德曼标度方法，它对联列资料进行检验，看是否存在着一系列争端，并根据他们在每一争端上的投票，理想地把他们置于这个系列的相当地位上。对所有争端投赞成票的议员将归于尺度的一端，对所有争端投反对票的议员归于另一端，赞成一些争端也反对另一些争端的议员将被置于尺度的中间位置。如果在资料中可以辨认出这类尺度，像艾德洛特教授在他对 19 世纪40年代英国议会的研究中所做的那样①，那么它可以为下议员的投票行为，以及，回到本节的主题，对一个争端的投票和对另一个争端的投票之间的关系，提供有价值的信息。

然而除了古德曼尺度这类方法，对定名资料关系形式的辨认很大程度上取决于资料及所研究的历史问题的性质。参考书目中列举的一些研究，显示了在这个问题上可以采用不同的方法。

另一方面，当资料为区间类型时，可利用的统计方法要多得多，我们可以试图用几种方法相当有把握地回答本节中早先提出的有关两个变量之间关系的形式的问题。对于区间资料，运用相关系数使我们能够确定两个变量之间是否存在着一种关系，而相关系数的符号告诉我们这一关系为正相关还是反相关。例如，从商船的吨位和船员人数的资料计算 R，显

① W. O. 艾德洛特:同前书，第134—163页。

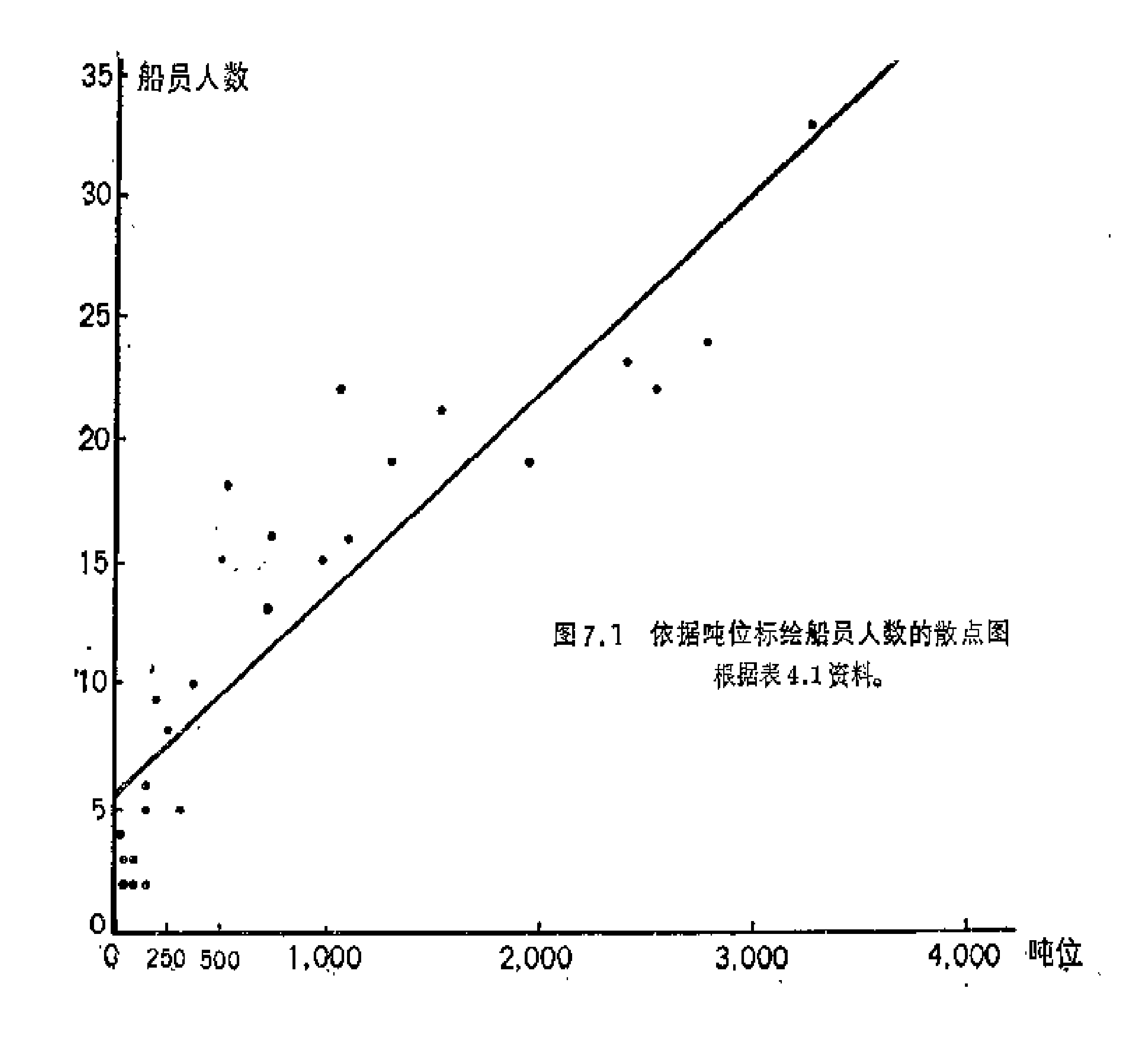


图 7.1　依据吨位标绘船员人数的散点图

根据表 4.1 资料。

示两个变量之间存在着一种关系，而且是一个相当强的正相关关系($R=+0.9093$)：当吨位增加时，船员人数也随之增加。

然而，相关系数及其符号，都没有明确地向我们提供很多有关变量之间关系的确切形式的信息。例如，它们并没有告诉我们一艘商船的载货量要增加多少吨位才需要一名额外的水手来帮助操纵此商船。可是，这类信息是有价值的；若我们对航运业中商船规模的增加对雇用海员的影响感兴趣，我们就需要这一信息。与此类似，有关1870—1914年间英国工业的主要争论之一涉及人均产出增加还是减少；有关一艘特定规模的商船需要多少船员的信息与这些资料相关，当这些资料可用以与更早些年代的资料进行比较时更是如此。

我们需要了解的是吨位上升与船员人数也上升之间的关系。我们应能够回答诸如这样的问题："吨位需要增加多少才产生增添一名船员的需求？"图7.1表示的是以水平轴为吨位、垂直轴为船员人数的商船资料散点图。从图中可以看出，我们方才所问"吨位平均必须上升多少才产生对一名额外水手的需求？"，等于问："沿水平轴移动多远才能在垂直轴上上升一个单位？"可以记得，它很像当我们在时间数列资料中计算线性趋势时所考虑的问题："我们必须沿代表时间的水平轴移动多远才能在垂直轴上引起上升？"这一相似提示我们可以用拟合一条尽可能接近资料集各点的线这一同样的方法去回答关于吨位与船员人数之间关系的问题。

可以记得，我们是通过最小平方法来对时间数列资料拟合了一条线。这一方法同样适用于目前的问题：截距项a具有完全相同的意义，它是拟合线在水平轴的零点上面与垂直轴相交的一点。斜率b在这里代表倍数，人们只有在吨位中

的变化乘以这个倍数以后才能发现船员人数的相应变化，正如在时间数列的例子里，它代表为发现出口的相应变动、年份变化所必须相乘的那个倍数。而且，与时间数列的例子一样，最小平方法公式的应用使得我们可以写出等式

$$Y=a+bX$$

这里 a 为截距，b 为斜率，Y 为船员人数，X 为吨位数。

表 7.6 显示应用最小平方法公式对商船资料的计算，结果我们可以用数值填入式中的 a 和 b，并说明吨位与船员人数之间的关系用下面的等式来描述

$$Y=5.4481+0.0082X$$

此等式所描述并拟合于资料的这条直线，在图 7.1 上显示出来。

表7.6 根据表4.1和7.5的资料计算回归线

$\Sigma Y=320$ $\qquad\Sigma Y^2=5932$

$\Sigma X=22319$ $\qquad\Sigma X^2=42313977$

$\Sigma XY=470050$

$$b=\frac{N\Sigma XY-\Sigma X\Sigma Y}{N\Sigma X^2-(\Sigma X)^2}=\frac{11751250-7142080}{25(42313977)-498137761}=0.008235$$

$$a=\frac{\Sigma Y-b\Sigma X}{N}=\frac{320-183.7970}{25}=5.4481$$

回归线 $\quad Y=5.4481+0.0082X$

这个把一条直线拟合于一个资料集的过程被称为“把线性回归线 Y 拟合于 X”；或称“把 $\overline{Y}$ 回归于 X”。在上面的例子里，我们已把船员人数回归于吨位。我们只要把船员人数称为变量 X，把吨位称为变量 Y 就能完成把吨位回归于船员人数的另一过程。然而，这没有什么历史学意义。看来极有可能性的是 1907 年船东们先建造他们想拥有的一定吨位的船，

然后再寻找操纵它们的船员；看来极不可能的是先找到一批船员，然后再建造一条适当规模的船来雇佣他们。换句话说，从历史学上讲，看来船员人数很可能取决于吨位数而不是相反，因此我们应试拟合一条回归线以回答这样的问题："船员人数精确地在多大程度上取决于吨位数?"因此变量 Y，即这一案例中的船员人数，被称为"因变量"，而Y变量所依赖的X变量被称为"自变量"。我们根据自己的历史知识来选择哪一变量应被认为因变量，哪一变量为自变量。

通过把回归线Y拟合于X，我们刚才得到的回归等式$Y = 5.4481 + 0.0082X$ 告诉我们吨位与船员人数两个变量之间按平均数计算的关系。事实上，这条回归线是对关系的估计；它是根据我们得到的资料所能做出的最佳估计，但是我们必须承认它仅是一个估计而已，因为这条回归线并未恰恰穿过所有资料点，而只是尽可能地接近它们。因此，我们需要了解回归线在多大程度上接近资料点，也就等于了解对变量之间关系的这一估计有效到什么程度。

如果对资料点的"拟合"很接近，因此估计有效，我们将有信心地说："按平均数计算吨位每增加 1 吨即产生对 0.0082 名船员的需求"(或者更方便地说"按平均数计算吨位每增加 1000 吨产生对 8.2 名船员的需求")；如果拟合较差，那么我们就不会如此充满信心，并在描述商船规模的变化对劳动力的影响时较少把握。

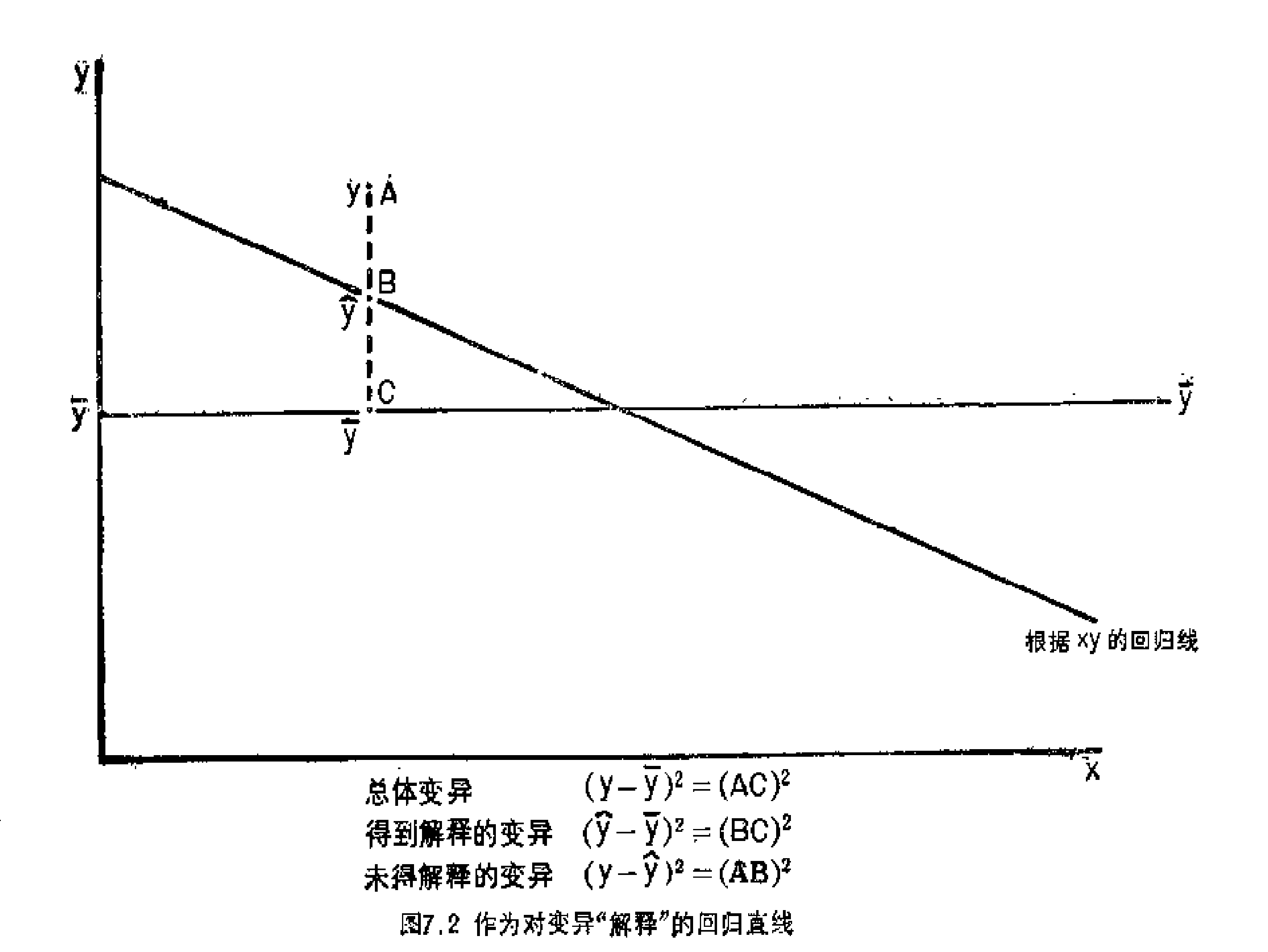
相关系数是对资料回归线的拟合有效度的一种测度方法。如果资料点都落在一条直线上，因而吨位越高船员人数也越高，则相关系数将为 1。随着资料点偏离回归线，相关系数将趋于零。正如我们在相关系数 $R = +0.9093$ 这个例子

里所看到的那样，变量之间的关系相当强，回归线对资料点的拟合也相当有效。

另一种考虑拟合有效度的方法是考虑X对Y回归线的计算在多大程度上有助于我们解释Y的变异。如果我们回想时间数列的例子，我们记得，在时间数列中我们计算了线性趋势，并得到了趋势值。于是我们同意这些趋势值可被认为是代表由时间流逝所决定或解释的那部分时间数列，因此我们从时间数列中减去趋势值；这一相减的结果被视为受其他因素，诸如周期性波动，所影响的那部分时间数列。当然，我们所谓"时间流逝"是指一整套变化条件，它在时间数列中用年份的流逝来表示。

类似地，我们可试将Y（船员人数）的变化分为两个部分，第一部分以吨位的变化来解释，第二部分归因于其他因素。用吨位解释的那部分将以回归线来表示（正如用时间解释的那部分以最小平方趋势表示），而用其他因素解释的那部分以资料点对回归线的离差来表示。我们把船员人数的变化或变异看作是一个围绕船员人数平均数的变异，正如我们以前用一个相对形式来表示它们那样。实质上，我们是认为船员人数由于若干原因而变异（围绕着它的平均数），其中之一为吨位的变化；我们想知道在多大程度上这一变异是由吨位变化引起的，在多大程度上是由于其他因素。

图 7.2 表明我们怎样做到这一点。对于每一个资料点，如图中的A点，离开平均数的距离被分为两个部分，第一部分是从平均数到回归线，而第二部分是从回归线到资料点。（观察这一过程的另一方法是考虑对回归线的了解能在多大程度上改进我们根据一个特定吨位值对船员人数值的推测。如我



总体变异 $(y-\bar{y})^2=(AC)^2$

得到解释的变异 $(\hat{y}-\bar{y})^2=(BC)^2$

未得解释的变异 $(y-\hat{y})^2=(AB)^2$

图7.2 作为对变异"解释"的回归直线

们对船员人数与吨位之间的关系一无所知，对任何吨位的船员人员的最佳推测将是船员人数的平均数。对回归线的了解可以使我们根据船员人数平均数与从回归线估计的船员人数之间的距离来改进这一点。衡量我们的推测在多大程度上被改进了，是对实际资料点的回归估计接近到了什么程度。

如果在这个基础上我们把平均数与回归线之间的变异视为船员人数由吨位影响所解释的那部分的变异，我们可以看出对回归线的拟合良好度（因而是我们估计吨位与船员人数之间关系的精确性）的衡量是这一"得到解释的变异"占船员人数有关平均数的全部变异的比例。

在图 7.2 中，我们仅对一个资料点说明了船员人数有关平均数的全部"得到解释的变异"与全部"未得解释的变异"之间的关系。为了用有关平均数的回归公式对所有资料点计算得到解释的变异，我们将所有的变异平方（以去掉正负变异相约的影响），并对所有资料点求和。于是其被称为测定系数的结果是

$$\frac{\Sigma(\hat{Y}-\overline{Y})^2}{\Sigma(Y-\overline{Y})^2}\quad \begin{array}{l}\text{（平均数与回归估计值之间的距离）}\\ \text{（平均数与资料值之间的距离）}\end{array}$$

这一测定系数的计算稍微有点费事，尽管回归线算出后再从事这项工作从数学角度上讲并不复杂。在这个例子里，测定系数为 0.8268，因此我们认为回归线对资料点拟合得很好，它可以解释Y变动的 0.8268 或 82.68%。换句话说，82.68%的船员人员变异可由吨位变动来解释。

在实践中我们并不需要直接计算测定系数，因为可以显示它等于相关系数R的平方。由于这个原而，测定系数通常被写作 R^2。因此，衡量回归公式的拟合有效度的最好方法为

相关系数的平方，因为它精确地告诉我们用根据X对Y的回归公式解释的Y的变异在多大比例上归因于X的影响。这一事实还可帮助我们判断上一节所讨论的两个区间尺度变量之间关系的强度。如果我们计算出两个变量之间的相关系数为0.9，那么应变量中81%的变异应由自变量的影响来解释；如果相关系数仅为0.6，则只有36%的变异得到解释。由于这个原因，当R小于0.7时，不应对两个变量之间的关系过分肯定，因为只有不到一半的变异可归因于这一关系的影响，而大部分是由其它因素所造成的。

最后，关于相关和回归分析需要强调四点。第一，我们在整个这一章里仅考虑了线性相关和线性回归——即资料点可以用一条直线来表示的情况。完全有可能两个变量之间存在着一种很强的非线性关系；如果是这样，线性相关系数R的计算将给出一个很低的相关值，而线性回归线对资料将给出一个很差的拟合。由于这个原因，在计算R和回归线之前，通常像图7.1那样将资料标绘在一个散点图上是明智的。只有当明显存在着直线关系时才能计算它们；如果一条曲线看来似乎能给出一个更好的拟合，高级统计教材将讨论适当的方法。

第二，我们只考虑了一种例子，在其中只有一个自变量被认为是影响了因变量。这里讲过的方法也可以扩展到具有两个或两个以上的自变量的例子中去。倘若如此，这些方法就是多元回归分析的方法，而不是这里所讨论的简单回归分析；同样，如需获得详细的指导，应参阅高级统计教材。

第三，相关与回归方法在历史研究中通常被用于根据有限的证据对两个变量之间的关系作一般性陈述。例如，根据我们的证据，即限于1907年25艘商船的资料，我们可能想对

船员人数与吨位的关系作出一般性陈述。由这样一种尝试所引起的问题将在下一章里进行讨论；这一章只关心我们手头所有的资料中的关系，而不是从这些资料中作出概括。

第四，我们必须再一次强调，只有当历史学家用一种清楚的理论把他试图描述其关系的变量联系起来时，相关和回归分析的方法，以及对定名和定序资料的类似方法才是有意义的并才应被应用。由于纯属偶然的原因，可能会出现拟合良好度很高的回归直线，因此较高的R^2值，但除非它们所要描述的关系能够以一种完全不同于统计方法的方式来理解和解释，不要信任它们。

7.4 含有时间数列资料的相关与回归

一个历史学家常会想去探索两个时间数列变量之间的关系。确实，它比起上一节所用的那类常被称为"横断面"的研究，其中没有任何时间成分的例子有大得多的可能性作为一个历史研究的项目。然而，应用那个例子的原因在于时间成分会使相关和回归分析复杂化，而这后两者只有当这些分析的基本原理被考虑了之后才能讨论。这里只能考虑这些复杂性中的一点，而任何真想对时间数列资料进行回归分析的历史学家应参考更高级的教材以获得更多的信息。

将回归方法应用于时间数列资料的主要复杂之处，在于两个线性趋势永远是绝对相关的——如果两趋势正向相同的方向移动是绝对正相关，否则是绝对负相关。这一点可以从图 7.3 中清楚地看出，两个线性趋势分别按时间标绘，以及在一个散点图中，那里所有的点都落在一条直线上，指示一个绝

对相关。因此，如果我们计算两个各自具有一线性趋势的变量之间的相关系数，R值将受到趋势存在的影响。假如我们对19世纪初期的英国的进出口之间的关系感兴趣，那么双方分别具有一个上升的时间趋势的事实将会使得相关系数相当高，而且是正数。如果我们感兴趣的是显示进出口双方都在上升，那么这不成问题。然而，更可能我们感兴趣的是进出口波动之间的关系；我们想知道进口中的一个上升是否导致出口的一个上升——如果是的，上升多少。如果我们的兴趣在于波动之间的关系，那么很清楚我们不希望相关和回归的估计受到线性趋势的影响。

所以对历史学家来说，了解时间趋势是否影响着他的结果是重要的。有时观察一个图表或散点图，以这样方式起作用的时间影响是很明显的，但有时这种证据可以是很模糊的。因此，利用某些测度可能扰动时间影响的方法是有意义的，为此许多历史学家采用一种被称为德宾—沃森检验的检验法，它导致计算一个叫作d^*的值。为了解这一也被称为自相关的检验方法，我们必须回到上一节所描述并在图7.2中画出的方法，根据这种方法应变量中的变异被认为由两部分组成：一部分由自变量的变异“解释”，而另一部分为“未得解释”的变异。对于每一个资料点，应变量的值与所有值平均数之差被类似地分割成回归线与平均数之差(图7.2中的BC)以及余项(图7.2中的AB)；这个余项通常被称为“残差”，而这些资料点的剩余集被称为“回归公式残差”。回归分析的一个重要假定是各连续残差之间彼此“不相干”；统计学上的不相干性的概念将在第八章第(2)节里详细进行解释，但大致讲，它意味着一个残差值不会影响次一个残差值。否则就违反了回归

的这一基本假定，自相关被说成存在，并且在回归方程

$$Y=a+bX$$

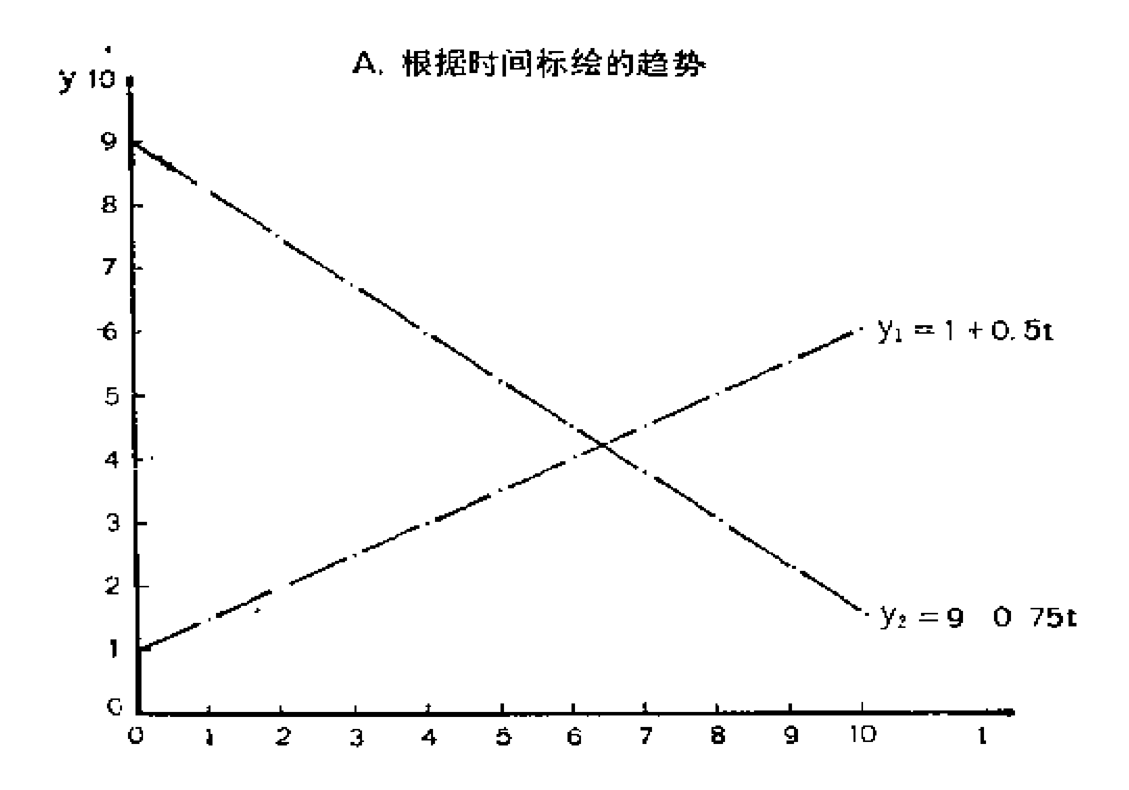
中的回归系数 a 和 b 以及测定系数 R^2 可能受到影响，致使它们不能真正表示 Y 与 X 之间的关系。

两个时间数列中的时间影响很可能使当一个时间数列对另一个时间数列求回归时连续的时期（例如连接年份）的资料点很可能接近在一起，其回归线残差因而很可能是相关的。由时间影响产生的自相关，对历史学家来说是一个普遍的问题，因此每当对两个时间数列求回归时应进行德宾—沃森检验。d^* 的计算十分简单，若取 e_t（这里 t 为时期）为残差，则

$$d^*=\frac{\sum_{t=2}^{n}(e_t-e_{t-1})^2}{\sum_{t=1}^{n}e_t^2}$$

对 d^* 值的解释通常要求参照一个图表集。作为一种粗略的指导，近似于 2 的 d^* 值表明不存在自相关，而接近于零或 4 的 d^* 值表明存在着正自相关或负自相关。使用这一检验方法的历史学家应报告 d^* 值及对它的解释。例如，N. 冯腾泽尔曼博士对 19 世纪初期英国进口与出口之间关系的研究，以对非欧洲国家（“其他出口”—X_0）的出口作为因变量①。在一项回归分析中包括谷物的进口为自变量或解释变量，而在另一项回归分析中不包括谷物的进口则为自变量或

① N. 冯腾泽尔曼:“论马修斯命题”(N. Von Tunzelmann, ‘On a thesis by Matthews’)，载《经济史评论》第 20 卷，1967 年，第 548—554 页。感谢冯腾泽尔曼博士和《经济史评论》的编辑允许我从这篇文章中翻印图表 7.4—7.6。



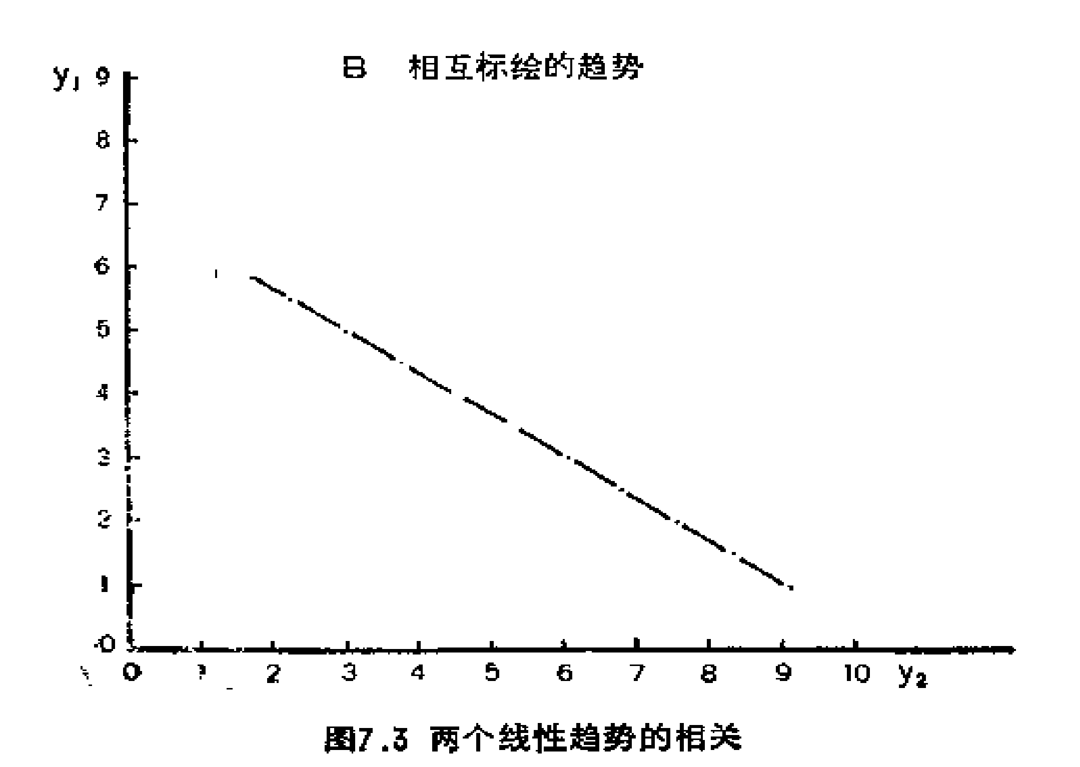


图7.3 两个线性趋势的相关

解释变量；在这两项分析中，由于兴趣都在于两个数列增长之间的关系，都用了自变量和因变量的对数。两个数列被画在图 7.4 中。在第一项分析中，两个数列之间存在着一个非常强的关系，如图所示并由 R^2 等于 0.991 所证实。然而，德宾—沃森统计量 d^* 为 3.55；这接近于 4，指出存在负自相关，这一点由"锯齿"型图解所证实。如图 7.5 所示，当 X 的实际值与估计值相比较时就会看到这种"锯齿"型图表。在第二项分析中，R^2 等于 0.965，d^* 为 0.85。这指出正自相关，而残差值的型式如图 7.6 所示；大多数初期值位于回归线的上方，后期值则位于回归线的下方，显示出时间对估计关系的影响。这一点也可以通过图解回归公式对时间的剩余值来加以证实。

然而，由于存在着使资料转换以消除一个扰动时间趋势的种种方法，时间影响并没有使分析变得不可行。第六章里描述的可以做到这点的一种方法，是从每一变量中算出趋势值，再从每一资料值中减去趋势值；这样就建立起一个于是可

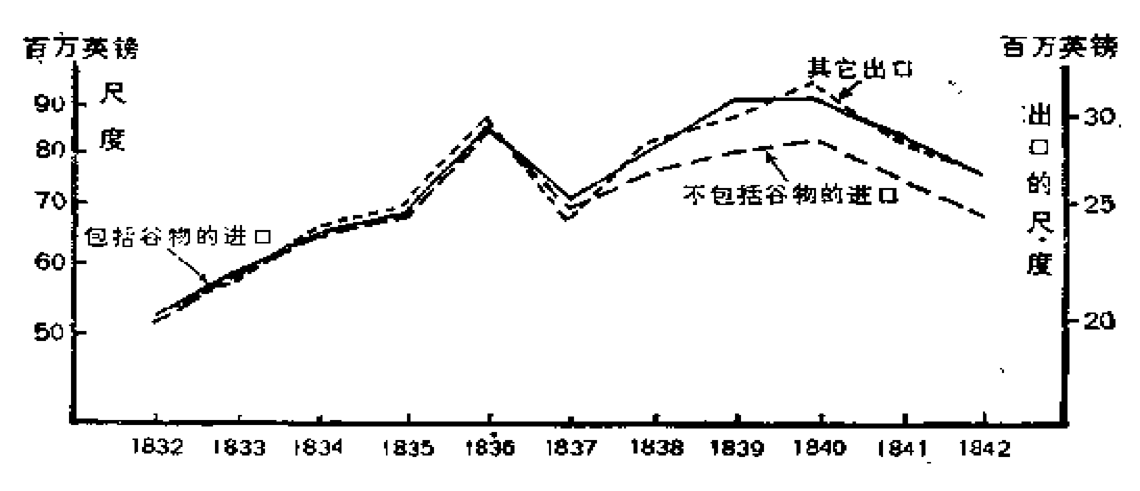


图 7.4　包括和不包括谷物的进口，及其他出口
资料来源：N. 冯腾泽尔曼。

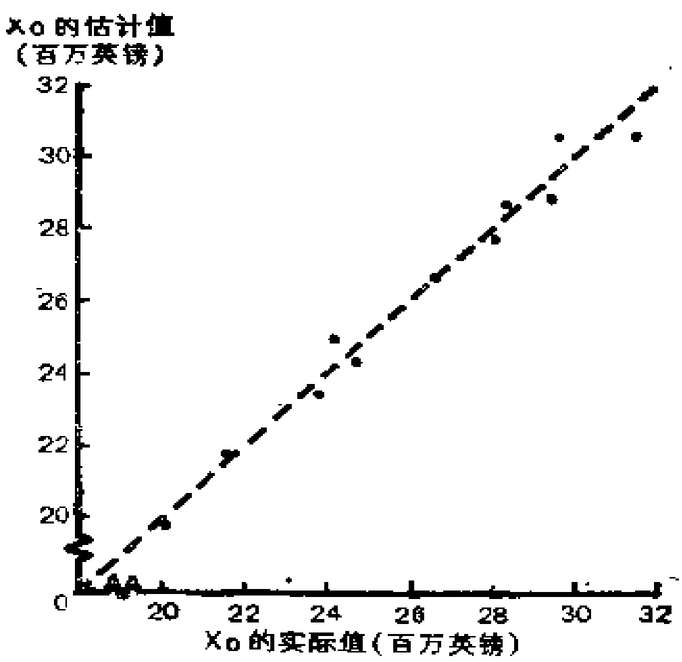


图7.5　其他出口的实际值和从其他出口依包括谷物进口的回归得到的估计值

资料来源：冯腾泽尔曼。

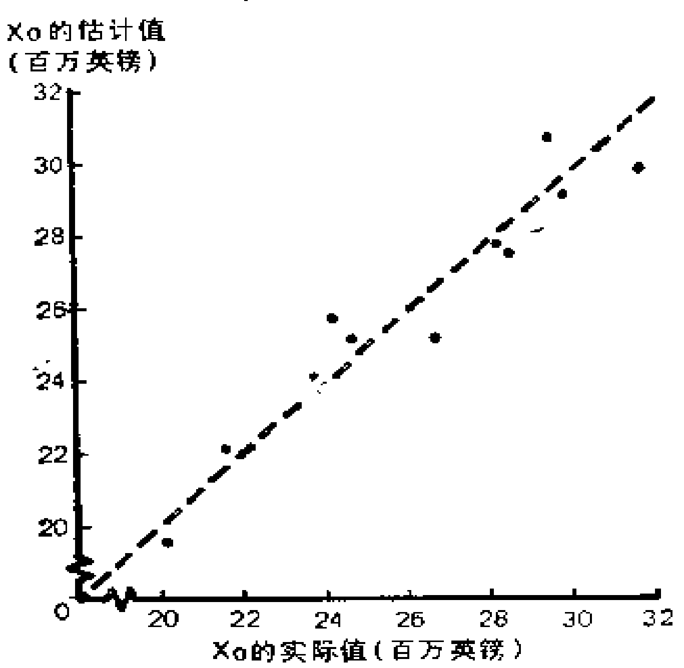
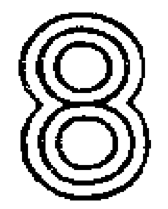


图7.6　其他出口的实际值和从依不包括谷物的进口的其他出口回归得到的估计值

资料来源：冯腾泽尔曼。

用于回归分析的“去除了趋势的”数列。类似地，通过从资料中消除一个强烈的周期性影响也可以产生一个新的数列。当我们的兴趣在于两个时间数列中逐年变化之间的关系（如图7.4中的逐年变化）时，另一种方法尤为适用，即从前一年值中减去每个每年值而为每一数列建立一新的数列。这种数列被称为一个阶差分数列。冯腾泽尔曼博士应用了一种稍微复杂的方法，并根据包括谷物的进口数列得到一个出口的回归及$R^2=0.993$，$d^*=2.04$；对于不包括谷物的进口，$R^2=0.882$，$d^*=1.80$。因此在这两个事例中，资料数列的转换排除了任何真正的自相关，但并没有严重地影响R^2度量的数列之间关系的强度。因此，自相关的存在并非时间数列的回归分析的一个难以逾越的障碍，但历史学家应意识到它可能带来的影响以及德宾—沃森统计的用法和意义。



资料缺失的问题

每个历史学家，当他研究一个历史问题时，都会在心中构成一幅他所要得到的，使他能解决其问题的证据的图景。我们可以认为这一理想是一个资料集，有了它就能为历史学家所面临的不论什么问题提供一个完整的答案。根据这一思路，每一历史问题会有自己的理想资料集，虽则很明显一个资料集会与另一个资料集相重叠，正如一个历史问题会与另一个历史问题相重叠一样。我们可以将每一个理想的资料集看作是由一组个案所组成；每一个案则由可以区别各个案的一组变量特征所组成，所有这些构成了一个资料矩阵。

在某些历史学问题中，要明确这类理想的资料矩阵将包括哪些内容是很容易的。假定我们对18世纪某一教区的人口史感兴趣；作为建立一个人口变化年表的最低需要，我们的理想资料矩阵应包括有关这一世纪本教区每一居民的出生、婚姻、死亡的资料。如果我们对探讨人口变化的原因感兴趣，可能还想要将诸如职业、收入等变量补充进资料集。对于其他一些历史学问题，对理想资料集的界定或许不那么容易，但毫无疑问可以想象这样的资料集是存在的。

在理想资料矩阵中，应能得到所有个案和每一个案所有变量的信息。我们应有足够的而且刚好足够的，使我们能回答所面临的问题的信息。早先在第二章中讨论过，我们将在收集证据的过程中，试图填充资料矩阵的内容，为每一个案及每一个案中的每一变量提供证据；这样在收集资料终结时我们将得到一个与我们的理想矩阵规模相同的真实资料矩阵；在此基础上我们开始对资料的分析。

不幸的是，以大多数历史学问题来说，我们能够收集到的真实资料不会确切地或完全与理想矩阵相符合。历史学家最通常的抱怨是“要是我们能更多地了解这个或那个该多好啊”，而紧接着这个抱怨的是。“我得到的资料太多了，真不知道该怎么处理它们！”换句话说，真实资料矩阵几乎肯定会在某种程度上与历史学家开始考虑他的问题时有意无意建立起来的理想资料矩阵不同。

真实资料矩阵可能在很多方面与理想矩阵不同。我们可以将真实矩阵对理想矩阵的歧异归纳为4种类型，认识到任何真实资料矩阵都可能表现出一种，几种或所有对理想矩阵的歧异。这4种是：

(1) 信息太多。

(2) 缺失有关一个或更多的完整个案的信息。

(3) 有关一个或更多变量的信息完全缺失。

(4) 缺失有关某些个案中某些变量特征的信息，但没有一个个案或变量的信息完全缺失。

奇怪的是，第一种可能与理想矩阵相歧异的是资料并不缺失；毋宁说问题在于历史学家的资料太多了，以致他不能有效地应用它们。我们在此考虑它是因为这种情况表示真实资

料集与理想资料集的一种歧异，并因为可能克服这一缺陷的方法与讨论缺失材料的情况有关。

8.1 信息太多：变量的选择

虽然应用计算机和电子计算器可使历史学家分析大量和复杂的材料，历史学家仍面临某些材料太多的情况。缺少电子设备帮助的情况仍然普遍，而且有些历史资料集大得甚至使用计算机其有效地处理资料的任务仍然十分巨大。在这种情况下，如果历史学家想要继续他的研究，就必须在现有的证据中作出选择，并在此选择的基础上作出自己的结论。换言之，他从自己的真实资料集中挑选个案和变量直至填充了他的理想资料集。他所面临的问题是确立作出选择所依据的原则。

我们将首先考虑对变量的选择问题，并以一种愈益被应用于历史探究的方法，即"集体传记"为例子；这个名词意为尽可能多地积累有关参与某些政治或经济活动的男女们的传记性信息。这种方法可与只考虑少数起主导作用的人物的较为传统的方法相对比。例如，艾德洛特教授收集了有关1841年议会所有议员的资料，并通过这些资料来研究议会的活动，而不是按照传统方法仅研究政党领袖皮尔·格雷厄姆、罗素·本廷克、迪斯累里及其他人的活动①。拉布教授并没有仅研究17世纪巨大的英国海外贸易公司的主要人物，而是收集了

① W. O. 艾德洛特："19世纪40年代英国下议院中的投票型式"（W. O. Aydelotte, 'Voting patterns in the British House of Commons in the 1840s'），载《社会与历史的比较研究》，第5卷(1963)，第134—164页。

有关向这些公司投资的所有人的资料。① 其他的研究还包括1789年大革命时期的法国人，② 20世纪30年代德国纳粹的支持者，③ 英国的实业家，④ 而“集体传记”这个名词还应用于有关若干教区、城镇乃至国家的资料的收集工作中。

在这些研究中，有关的历史学家必须决定对所研究的主题应当收集什么资料，因而不应当收集什么资料。这类决定往往由于现有证据的性质而强加于历史学家；例如，拉布教授将自己的研究限于每一投资者的3个变量：他的社会地位，他在国会中的议员资格及他所投资的公司。诸如出生与死亡日期、所任职务等其他方面的资料则被作为识别特定的投资者的辅助工具加以收集和应用，但它们不用于对资料的分析，因为只能收集到拉布所认定的8683名投资者中一部分人在这方面的资料。由于同样的原因，艾德洛特教授不得不略去有关下议员财产和宗教方面的资料。⑤

次一步要做出的决定是不收集某一特定变量的资料，因为这一变量被认为与对资料所要提出的问题无关。例如，拉布教授并没有收集有关“这一时期3个非常重要的新兴事业：……沼泽地排水、造船和渔业”方面的资料，因为“它们将会使

---

① T. K. 拉布：《企业与帝国》(T. K. Rabb, Enterprise and Empire)，马萨诸塞州，剑桥：哈佛大学出版社，1967年。

② C. 蒂利：《买主》(C. Tilly, The Vend'ee)，马萨诸塞州，剑桥：哈佛大学出版社，1964年。

③ W. S. 艾伦：《纳粹夺取政权》(W. S. Allen, The Nazi Seizure of Power)，芝加哥：1965年。

④ C.埃里克森：《英国的工业家：钢铁业和针织业》(C. Erickson, British Industrialists, Steel and Hosiery)，剑桥：剑桥大学出版社，1959年。

⑤ W. O. 艾德洛特：《历史中的计量化》(W. O. Aydelotte, Quantification in History)，马萨诸塞，雷丁：爱迪生-韦斯利出版公司，1971年，第146页。

我离题太远”。① 拉布教授这样说一方面承认对有关上述 3 项事业投资的资料极有兴趣，但另一方面又表明它们与他想要回答的有关英国海外投资的有限的几个问题无关。一个特定资料集是否与所考查的主题有关，必须根据对这一主题的历史知识来决定。在其历史知识的基础上，历史学家所必须做的是构筑一种将他想要提出的问题与他试图收集的证据联系起来的理论。例如，拉布教授在心里已形成了一种英国海外投资的决定因子的理论，这种理论将社会地位，作为议会议员的资格和在其他冒险事业中的投资看作是重要的，而将对其他新兴项目的投资（如他提到的 3 项）则看作比较不重要。既然这些其他投资比较不重要，有关它们的资料就不需要收集。

这种将问题与证据联系起来的理论，通常被冠以“模型”之名。在构筑一个模型时，历史学家取资于自己的历史知识，以及自己对例如经济理论的知识来绘出一幅有关一个历史事件或过程中的决定因子的图画。在此基础上他就可以收集有关的证据并试图回答使他感兴趣的那些问题。这种模型可能十分简单，仅联系少数几个变量。如拉布教授有 3 个变量。模型也可以非常复杂，特别是在经济史中；如福格尔和恩格尔曼教授构筑的一个模型，使用了 12 个变量来描述 19 世纪美国铁工业的增长。② 但无论模型多么复杂，其价值在于它能精

① T. K. 拉布：前引书，第 164 页。

② R. W. 福格尔和 S. L. 恩格尔曼：“对19世纪工业扩张的一个解释模型：在美国生铁业中的应用”(R. W. Fogel and S. L. Engerman, ‘A model for the explanation of industrial expansion during the nineteenth century: with an application to the American iron industry’)，载《政治经济学杂志》，第 77 卷，第 306—328 页，1969 年。

确地说明历史学家提出的有关证据的理论及不同变量之间逻辑关系的理论。在研究的过程中，模型当然可以被修改；研究的目的确实常是提出一个更好的理论以解释某些历史过程。但是，只有当我们对自己的模型或理论有了一个清晰的概念时，我们才能有一个坚实的基础来决定在研究中一个特定变量应被包括，还是排除。

迄今为止我们已经讨论了由证据的可得性和历史学家建立起来的模型或理论所决定的变量的选择问题。第三种类型的选择可以基于排除那些不能对从其他变量中得到的信息增加任何有价值的额外信息的变量。在最简单的事例中，信息可能以 2 种不同的形式出现。例如，我们可以从一个资料来源得知一个男人已婚，而从另一个来源得知他结婚时 26 岁。因为后者包含了前者，所以同时使用这两个信息毫无意义。在较复杂的事例中，历史学家可能会发现一个变量完全可以用另一个变量来代表。例如，在对英国铁路的影响的研究中，霍克博士仅考虑了铁路对小麦运输的重要性，但没有研究对其他谷物运输的重要性。由于小麦是运输中最重要的粮食，从这项研究中得出的各个结论就不会因为包括了其他谷物而改变，所以其他谷物被排除了。① 当然，这种省略一个变量的选择与其他根据模型所进行的选择一样，必须由历史学家来证明其正确性；但根据重复信息毫无意义这个总的原则来说，这样做常是可以接受的。

一般说来，只有作为历史学家有意识决定的结果，变量才

① G. R.霍克：《1840—1870年铁路与英格兰和威尔士的经济增长》(G. R. Hawke, Railways and Economic Growth in England and Wales, 1840—1870)，剑桥：克拉瑞敦出版社，1970年，第192页。

能被省略。甚至由于证据缺乏而不得不省略一个变量时，历史学家也常应意识到因此在他的分析中缺失了某些东西。正如艾德洛特教授所写，"学者必须以他对所省略的东西……以及所包括的东西的了解为指导，……并且必须小心行事，避免作出虽则与他所引用的数字相一致，却歪曲了他不得不省略的证据的任何推论。"①

8.2 信息太多：个案的选择

虽然除了谨慎和诚实之外，我们不能为变量的省略规定出什么总的原则，但个案选择的理论发展得很好。像"抽样理论"那样，它构成了大多数统计教科书的基础。因此我们不打算很详细地讨论种种抽样方法，而仅仅简要指出它们的主要原理。这样做既可使历史学家较容易理解论述抽样的教材，他们想要作抽样时，又可为本章后面要谈到的有关缺失资料的问题提供一个背景。②

在抽样中，我们从资料中作个案的选择。我们要减少必须处理的资料的数量，而又不大大降低从资料中得出的结果的准确性。因此，我们的目标是，根据我们对所选出个案的研究而得出的结论应与假若我们能研究所有个案而得出的结论相一致。换句话说，我们要样本为我们对真实结果提供一个准

---

① W. O. 艾德洛特，前引书，第147页。

② 对抽样问题的极好的讨论，参见 R. S. 舍菲尔德："历史研究中的抽样"(R. S. Schofield, 'Sampling in historical research')，载 E. A. 里格利编：《十九世纪的社会》(E. A. Wrigley, ed. Nineteenth-Century Society)，剑桥：剑桥大学出版社，1972 年。

确的估计数。例如，如果我们对发现1907年商船的平均吨位数感兴趣，我们将试图选择一个其平均吨位数与所有商船的平均吨位数相同的商船样本。与此类似，我们可能对商船的较复杂的特征感兴趣，或许是以蒸汽为动力的比例，或是船员人数的平均数和标准差。我们的目标总是一样：找到一个使我们可以对我们感兴趣的特征作出一个准确估计的样本。

从总的资料集中选择或抽取个案的任何方法，都将提供一个具有某些样本资料的据以作出估计的资料集。例如，我们可以只选取前10个个案，或者用一枚大头针随意挑选，或者选取我们所听说过的每一艘商船或那些具有有趣名称的商船。作出选择之后，我们就可以计算，例如，我们样本中的商船平均吨位数，然后我们会得到所有商船的平均吨位数的估计。不幸的是，我们没有办法证明这一估计准确到什么程度；它可能很准确，也可能谬误百出，而我们无法知道它属于哪一种。因此，抽样理论的意义首先是为我们提供一种能使我们得出准确估计的个案选择方法，其次要使我们能估计出这种估计可能准确到什么程度。

抽样理论和方法是根据两个概念——其一为正态分布，其二为独立随机抽样——以及由这两个概念推导出的一些定理。我们将分别考虑正态分布和独立随机抽样这两个概念，然后例示它们怎样协助我们决定抽样方法。

正态分布是频数分布的一种特定形式。它们具有这样的特定性质，即在分布中一个固定比例的个案处于分布的平均数与任何给定的离平均数的距离之间，而离平均数的距离则用分布的标准差的倍数来表示。例如，在分布中68.26%的个案落在高于平均数的一个标准差与低于平均数的一个标准

差之间；95.46%的个案落在高于平均数的两个标准差与低于平均数的两个标准差之间。如果我们有一个平均数为175，标准差为 25 的正态分布的资料集（即其形状接近于一个正态分布的形式），我们就知道在这个分布中 68.26%的个案的值介于 150 与 200 之间，95.46%的个案的值介于 125 与 225之间。存在着无穷多的正态分布，每一分布对应一组平均数和标准差，但它们都具有这种性质。而且，每一种正态分布都可以被转换成称为正态分布的标准形式；这是一个平均数为零，标准差为 1 的正态分布。任何其他正态分布的值都可以通过下面的公式转换成这一标准形式的值

$$Z=\frac{X-\overline{X}}{S}$$

这里 X 为原始正态分布的每一个值，$\overline{X}$ 和 S 为它的平均数和标准差，Z 为正态分布标准形式的每一个值。

图 8.1 为正态分布的标准形式，以及落在离分布的平均数的特定距离之间的个案数目。

抽样理论的第二个基础是独立随机抽样。随机样本意味着选取一个个案样本使每一个个案都有同等的被选为样本的

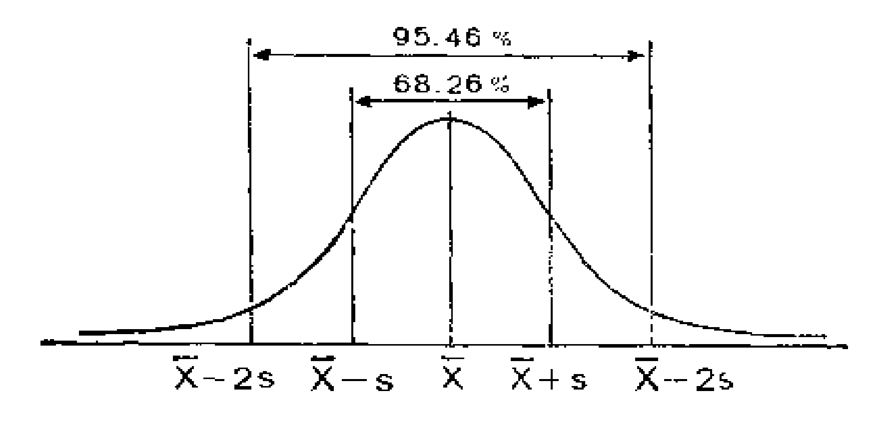


图 8.1　正态分布的标准形式

一部分的机会，而每一个个案组合都有同等的被选中机会。"独立"意味着将一个个案选为样本不应影响任何其它个案出现在样本中的机会。应该强调指出，只有当这些条件都得到满足时才能说样本是随机的；一个随机样本并不是从历史的意外事故中侥幸残留下来的东西，也不是由偶然投合我们兴味的个案所组成。在实践中，我们选取一个随机样本时通过利用随机数字表来保证上述条件得到满足。这些随机数字表构筑得使任何一个数字出现在表中任何一点或数字的任何组合出现在表中的机会都相等。因而随机数字表中排列的数字满足随机抽样的条件。表 8.1 所示为这种随机数字表的一小部分；完整的随机数字表以书的形式出版，也收编在统计学图表的书中。此外，许多电子计算器也可以产生随机数字。

表 8.1　摘自一张随机数字表

75	56	97	88	00	88	83	55	44	86
23	79	34	87	63	90	82	29	70	22
94	68	81	61	27	56	19	68	00	91
18	28	82	74	37	49	63	22	40	41
13	19	27	22	94	07	47	74	46	06

设想我们想要运用随机数字表从《末日裁判书》列举的埃塞克斯郡采邑中(见表 3.1) 选取一个随机样本。表 3.1 列举了 50 个采邑，我们打算从中选取 10 个随机样本。

我们首先为 50 个采邑编号，第一个采邑里特尔为1，未特别指明的采邑为50。这50个数字被称为我们进行抽样的"总体"。然后我们可以从随机数字表中任何一点开始；比如让我们从第 3 列最上面的数字 97 开始。它大于 50，因此我们不能用它。所以我们选取紧靠下面的数字 34 (同样我们也可水平

移动取88)。34小于50,因此我们选取第三十四个采邑埃尔森哈姆为样本中的第一个采邑。沿此列继续往下得到81和82,我们放弃不用,然后得到27;表中第二十七个采邑韦瑟斯菲尔特于是成为样本中第二个采邑。现在我们就可以移到表的其他列或其他行。假定我们只移到第4列的顶端,继续选取会发现22,因此取第二十二个采邑。再移到第5列继续选取会发现27。现在我们已经为样本选取了第二十七个采邑;如果再选取它,那么在样本中一个采邑将出现两次,这将是一种浪费,因为在样本中会重复信息。所以我们可以略去这个数字(虽然并不是必须要这样做)并继续移动,选取37,47,7,19,29(略去22),44,40(再略去22),和41作为我们的10个样本。如果对一较大总体(比如说3000个采邑)进行抽样,我们则将只取四位数的数字,并如前进行。

我们已经叙述了正态分布和随机抽样,现在我们就可以继续说明它们怎样与抽样理论相关。假如我们从一总体中反复地抽取N个个案的随机样本;这一总体有一平均数,用符号μ(读作mu)表示,其标准差用符号σ(读作sigma)表示。我们每取一组随机样本即计算该样本的平均数;随着我们所取的样本越来越多,这些平均数本身即构成一个有自己的平均数和标准差的频数分布。可以看出,当所提供的N(样本容量)足够大时(在实践中大于100),这一样本平均数的频数分布(抽样分布)将是一个正态分布。而且,抽样分布的平均数将等于总体平均数μ,而抽样分布的标准差(称为"标准误差")将等于$\sigma/\sqrt{N}$,即总体的标准差除以N的平方根。只要所提供的样本容量足够大,不管总体本身是否为正态分布上述这点都是正确的。

这些特征适用于抽样分布，而并不适用于单个的随机样本上。然而，由于样本分布是正态的，我们便知道分布中68.26%的个案将落在该分布的平均数和正负标准差的区间里。构成样本分布的个案是单个随机样本的平均数，因此我们可以说68.26%的随机样本将在这个范围里求平均数。我们可以把这句话换个方式说，在100个样本中的68.26个样本将在这一范围内求平均数，或者相当于说任何随机样本将有68.26%的机遇在这一范围内求平均数。类似地，我们可以说，任何随机样本有95.46%的机遇在总体平均数的两个标准差之间求平均数。

正是这些事实为应用抽样方法提供了正当的理由。我们知道，假如取一个相当大的随机样本（在实践中超过100个个案），随机样本的平均数将有很好和精确确定的机遇接近样本所从抽取的总体的平均数。即使样本较少，也可以精确地确定机遇。因此样本平均数是对总体平均数的一个良好估计。此外，如果我们知道了总体的平均数μ和标准差σ，我们即知道有68.26%的机遇确保来自总体、规模为N的一组随机样本的平均数将在$\mu\pm\sigma/\sqrt{N}$的范围里取值。这种了解在历史学问题中极为重要。历史学家愈来愈多地应用由19世纪人口普查中普查员所收集的资料，并从普查员的登记簿中抽取样本。这些登记簿中的许多材料被用于已发表的人口普查报告之中，诸如某一特定地区人们年龄的平均数和标准差这类资料都可以从中得到。因此，历史学家可以从普查员登记簿中抽取一个随机样本，了解100个样本中有95.46个样本的平均数应介于总体平均数加2个标准差及总体平均数减2个标准差的范围之间。若结果不如此，那他马上就会意识到要么他

在抽样时犯了一些错误，要么他的样本不幸正是100个样本中那4.54个样本之一，其样本平均数处于这一范围之外。若他的样本处于这一范围之中，那他可以相当肯定一切无误，并运用他的样本推知更多有关他所抽样的总体的信息（当然，也有可能他在抽样时犯了错误，但是样本平均数仍然落在正常的范围之内；因此，一个样本平均数，落在这一范围内，未必能保证不是一种坏的抽样方法）。

然而，在大多数历史研究中，总体的平均数和标准差是未知的。确实，提供对它们的估计常是样本的目的。在这种情况下，我们知道样本平均数（100个样本中有95.46个样本）应处于$\mu \pm 2\sigma/\sqrt{N}$的范围之内，但是由于我们不知道总体的标准差σ，所以也就无法知道这一范围是多少。我们所得到的唯一信息来自随机样本，因此我们必须利用这一信息帮助我们估计$\sigma/\sqrt{N}$。可以看出，对$\sigma/\sqrt{N}$的最好估计为$S/\sqrt{(N-1)}$，这里S为样本的标准差，N是样本的个案数，而且抽样是随机的。因此，当总体平均数和标准差为未知时，我们首先从总体中抽取规模为N的随机样本，并计算样本平均数和样本标准差。运用方才表述过的定理，我们可以说在所有样本中有95.46%的样本的平均数将落在$\pm 2[S/\sqrt{(N-1)}]$的范围内——即取自于总体平均数的样本分布的2个标准差（或标准误）。通过这种方法，我们利用样本对我们感兴趣的总体特征作出估计，它将是最可能好的估计。

我们可以通过一个简单的假设例子来说明应用抽样方法来提供对总体特征的估计。假定我们想要通过样本平均数对一个特定城镇妇女初婚年龄的平均数作出估计。我们从教区登记簿中收集有关结婚年龄的资料，然后取一个有100个妇女

的随机样本。初婚年龄平均数$\overline{X}$等于27，标准差 s 等于22年。根据上面讲过的定理我们知道，样本平均数是对总体平均数的最佳估计，并且在100个样本中有 95.46 个样本的平均数将落在 $\mu\pm2\sigma/\sqrt{N}$范围内。这等于说 100 个样本中 95.46 个样本的总体平均数将大致在 $\overline{X}\pm2\sigma/\sqrt{N}$的范围内。由于我们不知道总体标准差 σ，我们运用 $S/\sqrt{(N-1)}$ 作为$\sigma/\sqrt{N}$的估计值。这样，对这个例子，我们知道，100 个样本中 有95.46个样本的总体平均数将在以下的范围内

$$\overline{X}\pm2\frac{s}{\sqrt{(N-1)}}=27\pm2\frac{(2.2)}{\sqrt{(100-1)}}$$
$$=27\pm0.4422$$

因此，总体平均数将介于26.5578和27.4422之间。围绕平均数的值域±0.4422，被称为"95%的置信区间"，因为我们可以说有95%的把握总体平均数将落在这一区域内。

迄今我们已讨论了在对样本所从抽取的总体的特征作出估计时，抽样理论的应用。此外，我们可以应用样本理论帮助检验有关样本结果的假设。一个历史学家常有兴趣去了解他所研究的主题中的某一特征在两个时期之间是否发生了变化。应用我们的假设例子，我们可能对发现在一个世纪过程中妇女的平均结婚年龄是否变化这一问题感兴趣。由于一项婚姻可能生育的儿童人数与妻子的结婚年龄有关，在了解人口活动时知道大多数妇女的结婚年龄是重要的。因为我们所研究的婚姻数目十分多，我们需要应用样本。有关某世纪初100个婚姻的第一个样本表明，结婚年龄的平均数是27，其标准差为2.2岁，而有关100年后100个婚姻的第2个样本则表明，结婚年龄平均数为26.5岁，其标准差为1.6岁。初看起来，似乎

一个世纪以后结婚年龄平均数下降了0.5岁。但是我们必须记住这些只是样本结果，给我们的仅是对总体结果的估计；再则，由于我们是在抽样，必须时时记住存在着样本可能不会给我们一个对总体结果非常准确的估计的风险。例如，我们可以设想第一个样本过高估计了结婚年龄平均数，而第二个样本过低估计了结婚年龄平均数；在每个样本在估计中只需有0.25岁的误差，就可以使0.5岁的结婚年龄平均数的表面变化丧失。因此我们需要找到某种方法去区分作为抽样过程的结果而出现的差异，以及存在于总体的结婚年龄平均数中的真实差异。这两种可能性可对比如下：

1. 总体平均数之间没有相差，但作为抽样结果的样本平均数之间存在着相差。
2. 总体平均数之间存在着相差，它反映在样本平均数之间的相差中。

为在这两种可能性之间作出判断，我们利用一种称为平均数相差检验的检验方法，它建立在抽样的另一项定理和正态分布的基础上。这项定理说明，如果我们从2个总体中取具有较大规模的大量的独立样本，并计算每一对样本平均数之间的相差，那么这些相差的抽样分布本身就是一个正态分布。它的平均数将等于2个总体平均数之间的相差，而它的标准差（标准误）将是

$$\sqrt{\left(\frac{\sigma_1^2}{N_1}+\frac{\sigma_2^2}{N_2}\right)}$$

因此，我们可以利用正态分布的特征，而可以说，例如，两个样本平均数之间的相差将有95.46%的机会落在总体平均数之间的相差

$$\pm 2\sqrt{\left(\frac{\sigma_1^2}{N_1}+\frac{\sigma_2^2}{N_2}\right)}$$

的范围之内。

运用平均数相差检验，我们要探讨的第一种可能性是在两个总体平均数之间没有差异。即如果我们将 μ_2（第 1 个总体的平均数）从μ_1（第2个总体的平均数）中减去，结果将为零。倘若如此，样本平均数之间的任何非零的相差都将是抽样的随机结果；再则，只有 4.54%（100－95.46）的样本可能具有大于 2 个非零的总体标准差的平均数之间的差异。因此，平均数相差检验的逻辑性如下：计算两个样本平均数之间的相差，再除以标准误

$$\sqrt{\left(\frac{\sigma_1^2}{N_1}+\frac{\sigma_2^2}{N_2}\right)}$$

假如结果大于＋2或小于－2，那么如果总体平均数真相等，只有4.54%的机遇这一样本平均数的相差将会是抽样随机的结果。因此，若结果大于＋2或小于－2，要么是我们在抽样中非常不凑巧，要么是总体平均数不相等。若它们不相等，它们肯定有差别，而因此我们可以得出这样的结论：在样本所从抽取的总体中，结婚年龄存在着差异。

在实践中，我们用一个公式来计算 z——平均数的相差除以合并的标准误差。由于我们不知道总体的标准差，我们利用样本的标准差作为估计值。在平均数相差检验中求z的公式为

$$z=\frac{\overline{X}_1-\overline{X}_2}{\sqrt{\left(\frac{s_1^2}{N_1-1}+\frac{s_2^2}{N_2-1}\right)}}$$

对于我们用的假设例子，$\overline{X}_1=27$，$s_1=2.2$，$\overline{X}_2=26.5$，$s_2=1.6$，$N_1=N_2=100$，这样

$$z=\frac{27-26.5}{\sqrt{\left(\frac{2.2^2}{99}+\frac{1.6^2}{99}\right)}}=1.829$$

由于 z 为1.829，我们知道存在着一个大于4.54%的机遇——样本平均数之间的相差为使用样本的结果，而总体平均数之间则没有差异。使用了正态分布表，我们可以更确切地说，尽管总体平均数相同，但存在着一个6.73%的机遇——所观察到的相差为抽样随机的结果。

假设我们现在取另一个100年以后的100项婚姻为样本，并发现这第三个样本的结婚年龄平均数为24，其标准差为2.1岁。将平均数相差检验应用于第二个和第三个样本，我们发现

$$z=\frac{26.5-24}{\sqrt{\left(\frac{1.6^2}{99}+\frac{2.1^2}{99}\right)}}=9.42$$

检验的结果 z 远远大于 +2，这表明样本平均数的差异是抽样随机的结果的可能性很小；这一可能性大大少于1%，但我们仍可以得出这样的结论，即在样本从以抽取的2个总体结婚年龄平均数之间存在着真实的相差。

对从同样的资料中的其他估计值可以作类似的检验方法，诸如比例或者回归系数。例如，若上一章所用的有关1907年商船的资料是通过一个抽样过程收集的，那么去检查这些结果是否由抽样过程所致而不是代表据以抽取资料的基本总体，将是非常明智的。在第七章里，我们推断出两个变量之间的关系为

$$\text{船员人数}=5.4481+0.0082\text{吨位}$$

表明这一关系是正的，即商船规模每增加1吨，船员人员平均

增加 0.0082 人。但根据样本证据，我们能否肯定也是如此，尤其是能否肯定我们的船员人数随吨位增加的结论；换句话说，就所有商船的总体而言，我们能否肯定吨位的系数不为零？

如同在平均数相差检验中那样，对一个特定的回归系数不为零的假设的检验是基于对回归系数的值与抽样分布的标准误差的比较，而抽样分布的标准误差得自于从一个其中真正的回归系数为零的总体中抽取大量样本。标准误差由以下公式估算

$$\sqrt{\frac{\Sigma(Y-Y')^2/N-2}{\Sigma(X-\bar{X})^2}}$$

这里Y'为从回归公式得到的 Y 的预测值。在商船资料的例子中，标准误差为0.00079；把0.0082(回归系数)除以0.00079我们得到值10.3797，它提示0.0082的样本结果不见得是产生于回归系数为零的总体。通过用计算出来的10.3797值与正态分布表相比较(或者，在这个例子中由于样本规模很小，可用另一种称为 t 分布的概率分布)以确定在概率分布中离平均数的标准差高于10.3797的值所占的比例，这是一种更为精确的检验方法；上述值的比例大大低于0.5%，而且我们可以有把握地断定样本不是从一个回归系数为零的总体中抽取的。

想要对从样本资料计算出来的回归系数（或者其他统计量）的可靠性进行估价的历史学家，需要知道标准误差的值，并通常用下面的形式来表示回归结果

$$Y = 5.4481 + 0.0082X$$
$$(0.00079)$$

括号中的项为标准误差。换一种方式，它可以表示为

$$Y = 5.4481 + 0.0082X$$
$$(10.3797)$$

这里括号中的项为 t 统计量，在前段中它是由回归系数除以标准误差算出的。

正如本节所示，不同的检验必须用于不同的样本结果，我们所做的不过是介绍了一些重要的抽样方法和概念。因此，任何想要从事抽样运算的历史学家，在开始他的研究之前必须更广泛地阅读有关抽样理论的书籍。

8.3 抽样结果的"显著性"

在讨论应用抽样方法估计总体特征及在讨论平均数相差检验中，我们描述了怎样可能计算样本结果成为总体结果的良好估计的问题。我们已用 68.26%的机遇或 95.46%的机遇这类话来表达这种可能性，或者如在最后一个例子中，我们说样本平均数的相差有6.73%的机遇是由于抽样随机引起的。一旦我们计算了这些机遇，我们就需要用它们来作出历史判断；例如，我们需要决定我们是否将认为总体的结婚年龄平均数中存在或不存在变化。我们知道样本相差由抽样过程产生的机遇为6.73%。但这一机遇是大还是小呢？

我们是否应接受一个机遇（一个6.73%的机遇）并说在两个总体的结婚年龄平均数中存在着一个真实的差别，这个决定必须是一种历史判断。统计学方法能够告诉我们机遇是多少（或者概率是多少），但是我们必须决定愿承担多大风险。我们的决定在某种程度上取决于特定结果对我们的探究所具有的重要意义；如果我们不过是偶然对结婚年龄平均数发生兴

趣，并且它究竟有没有变化对其他事物没有关系，那么我们或许愿意接受一个较大的风险。另一方面，如果我们的整个理论有赖于准确地知道结婚年龄中所发生的变化，那么我们大概只愿意接受一个很小的风险。通常在社会研究中所冒的风险水平为10%、5%、1%，常被称为10%，5%，1%的“显著性水平”。这一描述引起了给平均数相差检验这类检验以“显著性检验”的名称，并出现了“结果显著于 5%的水平上”之类的说法。这种说法只不过意味着为由抽样过程产生的结果的风险等于或小于 5%。意指同一事物所作的相等说法为“结果显著($P\leqslant0.05$)”，或“零位假设可在5%的水平上被抛弃”；零位假设通常是假设结果是由抽样的机遇所产生，因而人们抛弃它就是接受结果为准确反映了总体特征的假设。例如，在平均数相差检验中，零位假设为在总体平均数中不存在差别。

在上述或者类似的说法中，“显著的”这一名词只指在一特定的置信水平或显著性水平上，结果是肯定还是抛弃一个假设的问题。它与结果是否具有历史学意义这一点无关，虽然一个结果的统计学上的“显著性”在历史学家据以作出的结论中可能会成为一个因素。

显著性检验的使用不限于区间资料的样本，虽然我们讨论过的检验只适用于这类资料。如果只能得到定名或定序资料，可以使用其他一些检验，它们常被称为“非参数检验”，其中最常用的是以我们在推导列联系数时所用过的卡方统计量为基础的。这些检验方法的逻辑与我们已讲过的检验的逻辑非常相似，不过它们所利用的是其他形式的概率分布，而不是正态分布。

当考虑样本问题时，显著性检验很有价值，在对样本资料

进行任何估计或推论时，它们的使用是必要的。然而，它们容易被误用。首先，当，而且只有当资料按照诸如这一节前半部分所叙述的简单的样本随机抽样这种概率抽样方法收集时，它们才是适用的。如果样本不是一个概率样本，那么对它进行显著性检验从理论上讲毫无意义，并可能导致错误的结果。其次，我们所讨论的许多检验只有当其他限制性假设得到满足时才能应用：例如，平均数相差检验所假设的是区间资料，正态总体或者一个大的样本容量，以及独立的随机抽样。若没有满足这些条件，显著性检验同样毫无意义并会造成错误。

毋庸置疑，这些检验不应该用于非概率样本，亦不能用于违反了由这些检验所作出的任何假设的资料。然而，对于这些检验可在多大程度上用于根本不是样本资料这一点，统计学家与社会科学家有着不同的看法。例如，假定一个历史学家对研究1800—1900年牛津大学和剑桥大学的学生毕业以后10年所获得的平均收入感兴趣，以期发现牛津大学的毕业生以金钱而论是否比剑桥大学的毕业生更为成功。想象他设法发现了所有毕业生的收入，因而毫无缺失资料，并且他发现事实上牛津大学毕业生具有较高的收入。在这种情况下，许多社会科学家都很想进行平均数相差检验，并在最终给出结果时，以诸如“收入平均数之差在5%的水平上显著”的形式引述检验的结果。从算术角度上看这样做完全是可行的，但是难于了解这一检验以及它的结果的含义是什么。我们并没有一个毕业生的样本，而是全体毕业生。因此我们无法检验样本结果是否不同于总体结果；不言而喻，两者是相同的。在这种情况下，为了使显著性检验具有一定的意义，我们事实上必须假设存

在着某一假设的、较大的牛津—剑桥毕业生的总体，其中所有实际的牛津—剑桥毕业生组成一个简单的随机样本。然而，很难相信存在着假设的总体及随机抽样，因而显著性检验仍然没有意义。在收入平均数之间存在着差别，只能说到此而已。

当一个历史学家想要对一项特别研究进行概括时，如从一个工业城镇推论所有的类似城镇，从一个企业推论整个行业或从一个人推论一群人，常常会面临这种困难。在这种情况下，可以从算术上对一个城镇的两个特征，如公共事业的规模与支出的相关系数进行显著性检验。当发现在一个城镇中这种联系十分显著时，历史学家或许还会以从这个城镇概括其他城镇为快。但实际上，进行一项显著性检验并不能帮助他决定他是否能从一个城镇概括所有城镇；只有当一个城镇构成所有城镇中的一个简单随机样本时，它才会有所帮助。只有当历史学家能够证实的确如此，或他愿意作这样的假设时，显著性检验才有意义。

虽然抽样理论以及在实际中抽取样本有各种困难，历史学家应无犹豫地运用抽样方法而不应由于存在着大量的记录而放弃一项分析计划。由于能节省时间和费用，运用样本的优点是非常大的。这一点尤其如此，因为一个样本的准确性取决于样本容量本身，而不是它在整个总体中所占的比例。这可以从这一事实中看出来，如果抽取大的随机样本，从这样本作出的估计的准确性由σ/N这个量来决定，这里 σ 是总体标准差，N为样本的绝对容量。总体的规模并没有在这个量中出现，因此它与决定样本结果的准确性不相干。由于这个原因，取10%的样本而不是取总体其它比例的样本并没有什

么优点，问题在于样本的绝对容量。从 σ/N量中作出的一个更为重要的推论是，样本的准确性取决于样本容量的平方根而不是直接取决于样本容量。例如将样本容量加倍，它的精确度只增加$\sqrt{2}=1.4142$倍。与此类似，为了将样本估计的精确度加倍，我们需要将样本容量乘以4，因为$\sqrt{4}=2$。样本结果的准确性由样本容量决定，以及增加样本容量并不相等地增加它的准确性这两个事实，意味着有时有可能从一个较小的样本中导出完全可以接受的结果，而所需的努力与研究整个总体相比非常小。

8.4 资料太少：缺失资料的问题

“缺失资料问题”指在资料收集的过程中可积累的资料矩阵不能填满历史学家所想要填的理想资料矩阵的一切情况。我们在本章开始所区别分类的(2)、(3)、(4)类型都属缺失资料类型的例子。这类问题在历史研究中既比资料太多所引起的问题要常见得多，而且解决起来也困难得多。一般它们是由于记录受到破坏或者由于昔日官僚们未能保持我们想要得到的那类记录。第一类例子是，在英国史中有关洗礼、婚姻和丧葬的教区记录在有些教区被保存下来了，而在其他教区则没有。第二类例子是缺乏足够的19世纪以前的大多数国家人口普查资料。应该强调的是，有助于解决缺失资料问题的统计学理论很少，因此在解决这些问题时历史学家在很大程度上要依靠自己的想象和才智。

8.5 一个或更多的个案资料缺失

首先让我们看类型(2)，其中缺失某些完整个案的资料。在这种情况下，我们得到的是一个个案样本而不是一个概率样本(除非在极不可能的情况下)，在一个随机抽样过程后这些个案仍然幸存下来。由于有了样本资料，我们就可以利用它们来估计整个资料集的特征，正像如果整个资料集得以幸存下来我们将会估计的那样。然而，由于得到的不是概率样本，我们仍然无法确定这种估计准确到什么程度。当然，这并没有使这种估计变得毫无价值；它们自身可能是极为重要的结果。例如，在商业史中一个常见的特征就是所保存下来的信息大部分来自于那些兴旺发达和成功的商行；破了产的商行很少保存它们的记录。因而一个研究商业史的历史学家常面临着一些个案(商行)的资料缺失的情况。虽然这样，他所拥有的这些信息仍可以告知他许多有关成功商行的运转方式的情况，虽然从统计学上看这个信息不能用来估计这一行业中其他商行的行为，但它可能仍然是重要的。

在商业史的这个例子中，用一个简单随机样本的标准来评价，很清楚样本有两方面的欠缺。第一，个案的选择不是用随机手段做出的；第二，很明显样本具有倾向性，因为只有成功商行的信息才被保存下来。换句话说，这种样本不具有代表性。这产生了这样的问题，即一个似乎没有任何倾向性的样本，为了分析的目的，是否应被当作如同一个随机样本来对待。它将能使人们对从中得出的结果的准确性作出陈述，它还能为运用那些如同是从没有缺失资料的资料集中所得出

的结果这一作法提供合理性。例如，L.斯通教授在对16世纪英国贵族的研究中提出一个样本，它是由买卖双方任何一方姓名开头字母为S而被选中的采邑所组成的，并被看作如同是一个随机样本。他问道："有什么理由料想在这个事例中以字母S进行选择将会得到一个与真正的随机样本的结果大不相同的结果呢？……我有意避免使用字母J、O和M，因为它们可能在威尔士人、爱尔兰人和苏格兰人中产生不相称的数字。而在所有英国姓名中10%以上是由字母S开头，而且我看不出这个特定组会有什么独特之处。"①

斯通教授也清楚，按照严格的统计学理论这一方法并不能构成一个随机样本的抽取，而且同样从理论上也不能证明把一个非随机样本看作如同是一个随机样本是正当的，即使看来没有什么理由认为这一样本为何丝毫不具有代表性。然而在实践中，历史学家按照斯通教授所用的方法从事自己的研究似乎是正当的，只要，第一，他们知道自己在做什么，第二，他们清楚地向读者讲明他们在做什么。这一见解所产生的主要困难在于不同历史调查结果之间的可比性问题；例如，若我们像斯通那样用一种非随机方法对1535年英国采邑选择样本，然后再以另一种非随机方法对若干年后的采邑名单进行选择。在每一时期它们在所有权之间的不同是归因于真正的变化还是归因于两种非随机抽样方法之间的不同呢？对此人们只能说不知道，并且又为英国土地所有权变化的所有其它潜在原因之上增加了这个新的可能性。但是由于历史学家

① L.斯通："劳伦斯·斯通和采邑：反驳"（L. Stone, 'Lawrence Stone and the manors: rejoinder'），载《经济史评论》，第24卷（1971），第116页。

和读者都意识到了这一切，很难认为这一方法完全是错误的；我们将得到一些知识，纵然当更多的研究完成以后它必须被修正。

讲清楚已经偏离了严格的统计学方法的重要性怎么强调也不为过分。这种偏离在历史学研究中非常普遍。另一个例子是，人们曾做过一系列尝试，将从保险记录中发现的几个工厂的平均值乘以已知曾经存在过的工厂总数目，以估计工业革命时期棉纺厂的总值。严格来讲，这一过程包含了这样的假设，即那些具有已知值的工厂构成了一个所有工厂的随机样本，这不可能是真实的。虽然如此，只要清楚说明这点，只要不在这结论上作出过多的推论，这一方法是无可非议的。在实施这类方法而没有认识到或讲明所作出的是什么样的假设，这才是不正当的。

因此，我们对在其中个案资料缺失的类型(2)的讨论的结论必须是，由于缺少随机抽样，尽管严格说来这种情形常常近于无望，但在一些特殊的事例中运用违反统计学原理的方法也有可能得出近似的答案。对历史学家来说，首要的是更多地了解这些方法所必然带来的后果。

8.6 一个或更多的变量资料缺失

对一个或更多变量的资料完全漏失的情形，历史学家是十分熟悉的。例如，研究商业史的学者常对制造活动了解很多，但不知道所制造的商品的总值。政治史学家对他们所研究的政治家的经济状况通常知之甚少或完全不知，尽管他们可能了解其职业和家庭出身这类细节。在大多数事例中，对

一个变量的资料完全缺少是无法补救的。不存在据以估计变量值的信息，更说不上评价这一估计的准确性了。面对这种情况，历史学家只能寻找更多的材料，或者把他的研究限制在没有这些缺失资料的帮助也能回答的问题上。

当历史学家可以确定在他所拥有的变量值之间存在着某种逻辑的或统计学上的联系，从而他可以估计缺失的变量的值时，这种绝望的情况会出现例外。在一个琐细的事例中，一个经济史学家知道生产一件制成品的成本及其出售价格，但不直接知道所得利润，可以运用利润等于价格减成本这一知识来精确地估计缺失变量的值。在这种情况下，变量之间的关系是直截了当的，没有人会对这一过程提出反对意见。

若估计和被估计的变量之间的关系不是从一个简单的算术关系，而是从一个理论模型或从另一个历史时期或地点的证据中导出时，会发生更多的困难。运用这类方法试图估计一个缺失资料的值，尤其是被称为"新经济史"或"计量经济史"的显著标志——计量经济学是经济学中的这一分支，它运用统计学和数学的方法根据证据去检验经济理论。然而这种方法有其更广泛的应用，而"新"经济史学家主要在阐明这一常有意无意地为其他历史学家所使用的方法。例如，政治史学家常用诸如"自由派"、"保守派"或"法西斯主义者"这类描述来称呼人，虽则并不存在具有这些名称的政党，而且这些人本身也不会承认这些描述。在使用这些描述时，历史学家实质上所做的是表明他拥有对若干变量的信息，诸如对某些议案的态度，政治辩论中的行为，等等。他感到可以用这个信息使他能估计我们称之为"所研究的个人的政治信仰"这种缺失变量的值，并且，在称某人为"自由派"时，历史学家正在对一

个缺失变量的值作出估计。

为了说明所必须使用的方法的逻辑和一些可能遇到的危险，我们可以看两个例子，一个来自于政治史而另一个来自于经济史。第一个例子为艾德洛特教授为19世纪40年代的英国各届议会建立一系列政治态度所作的努力。通过对大量不同议案的表决的研究，艾德洛特教授认为他能够建立一个议案的尺度表，并可以根据他们对每一个议案的态度沿这一尺度表将议员分成等级。对所有议案投赞成票的议员被列在尺度表的一端，投反对票的列在另一端，然后艾德洛特教授就可以进而试图识别在政党忠诚、背景或意识形态中是什么压力把议员置于尺度表的特定点上。实质上，他正试图从现存资料的一组变量构成的政治行为中确定政治态度这个缺失的变量。对这一做法可能作许多批评，例如人们或许要问，艾德洛特是否完全排除了政党压力或政党恩惠的影响，当时的政治家是否把对一些特定议案的投票看得很重要或无关紧要。尽管有这些困难，艾德洛特的工作大大增加了我们对19世纪40年代的认识，而且在他在所有工作中都审慎地说明所用的方法，从而使批评家得以讨论他的方法。

估计方法在经济史中应用得最为明显。我们以迪恩女士和科尔教授有关英国经济增长，尤其是他们试图估计18世纪英国的谷物产量的工作为例。尽管有谷物进口和出口的数字可得，但没有任何国内粮食总产量的数字。迪恩和科尔的方法是，在1766年对每人每年所吃的谷物量作出估计，再乘以每十年期第一年的估算人口数。然后加上出口的粮食量，减去进口的粮食量，并把需要生产出来作为种子的那部分算进去，就得到谷物总产量的估计数。迪恩和科尔非常清楚地说明在这

一过程中可能出现的误差，及因而在以后的结果中的误差。他们的方法涉及以下假设：

(1) 当时对人口数目的估计是正确的。

(2) 对 1766 年谷物的平均消费的估计是正确的。

(3) 在 18 世纪里粮食的平均消费没有变化，尽管平均收入和粮食价格发生了变化。

(4) 准确地估计了 1766 年的粮食留存（即所需种子的数量），而且在整个 18 世纪期间这一比例没有变化。

所有这些假设为迪恩和科尔所卫护，并且是合理的，至少当计算的目的是要得出对谷物产量这个缺失变量的估计时是如此，这一估计是大致准确而不是非常精确。①

计量经济史中常用的填补缺失资料的另一种可能的方法是运用回归估计。例如，假设我们研究1907年的商船，并发现有关一艘商船的资料不完整；船员人数未被记录。如果我们已经从一个事先适当设计的样本得到了有关商船的资料（如上一章所用的资料），而且如果我们确信从回归估计所得的结果，那么可以运用回归公式来帮助我们估计缺失的船员人数。如果我们知道这艘船的吨位为 1600 吨，可以用 1600 替换回归公式 $Y=5.4481+0.0082X$ 中的X，并计算Y：

$$Y=5.4481+0.0082(1600)=18.5681$$

将小数点后面的数字四舍五入，我们就能够说根据现有资料

---

① 参见 P.迪恩和 W. A.科尔：《1688—1959年英国经济增长》(P. Deane and W. A. Cole, British Economic Growth 1688—1959)，剑桥：1964年，第62—68页。对这些假设的讨论见 N. F. R.克雷夫斯：‘18世纪的英国经济增长’(N. F. R. Crafts, ‘English economic growth in the eighteenth century’)，载《经济史评论》，第 89 卷（1976 年 5 月）。

我们对这艘船的船员人数的最佳估计是它有19名水手。当然，同样的资料也可以通过观察图 7.1 中所标绘的回归线更迅速地得到，即使它的精确性要稍微低一些。

从这些例子可以看出，对缺失变量值进行估计的可能性取决于将其他变量与漏缺变量联系起来的理论根据，并取决于现存资料的可信程度。再则，为构建他的论据，历史学家必须考虑对缺失变量进行的特定估计的重要性；如果这一估计对他的解释具有决定意义，他会希望对它们更可信赖而不是无足轻重。因此，在任何研究中，历史学家必须自己判断能否从事这类估计，而他的读者则必须判断他是否正确；为了后一个目的，历史学家必须清楚地说明作出估计的根据。

8.7 一个或更多个案中的一个或更多变量的资料缺失，而不是整个个案或变量的资料缺失

前两节讨论的方法也适用于第 3 种也是最常见的缺失资料的类型——缺失零散的资料。也就是可以用这一变量的其他值或其他个案来估计，或者运用这一个案中的其他变量值来估计缺失值。因此，对于资料矩阵而言，既可以垂直地(根据变量)也可以水平地(根据个案)作出估计，或者同时使用这两种方法作为检验。运用两种方法分别得到的两种估计之间的矛盾将启发估计方法可以改进和估计可以调和的手段。

对这最后一种类型来说，克服缺失资料这一困难的可能方法的范围比前2种类型较为广泛，但仍遭到同样的反对。第一，从其他个案在矩阵的垂直方面得到种种估计时，仍无法假设信息存在的这些个案可为矩阵中的所有个案构成一个随机

样本，因而估计可能具有倾向性的危险依然很强。第二，同一个案中的其他变量在矩阵的水平方面进行估计的可能性仍旧取决于对变量之间的关系所作的种种假定。应再一次指出，对估计方法的任何运用应尽可能清楚地加以说明，以便使这一方法能得到充分的批评和讨论。

9 计算器、计算机和历史资料

本书所阐明的所有统计学方法都可以用计算器来完成。况且，这类计算器的价格，以及因而它们所体现的统计能力的价格，近年来迅速下降，而且很可能继续下降。同时，由于计算机的广泛应用，可用于处理和分析历史资料的大得多的统计能力已可资利用，现在或不远的将来，许多国家里的学校、学院和大学都将拥有计算机。被称为“微处理机”革命的冲击是如此之猛烈，确实已难于在计算器和计算机之间作出区分，因为可编程计算器在性能和价格上可与微型和小型计算机相竞争，而从历史学家的许多实际应用上看，后者与大学里的大型计算机在性能方面不相上下。与这些计算“硬件”的发展相平行的是在“软件”中所取得的相似进展，各种程序和程序集使计算器和计算机可以更有效地被使用。

历史学家无论是在学校，大学甚或在家中，由于可以通过电视或电话使用计算机，因此面临着多得不可胜数的辅助手段来计算和分析他的材料。本章的目的是解释一些计算名词，消除它们的一些神秘，使历史学家能对为他解决历史问题提供最佳帮助的设备和方法作出明智的选择。

9.1 设备的选择：电子计算器

大致来讲，本书第一到第五章所描述的任何统计运算既可以用手工来完成(虽然这样做可能费时和麻烦)，也可以使用最简单的袖珍计算器来完成。为了进行第六章中的分析，可按这一章中若干表格所表述的方式使用一个简单的计算器①，但是许多稍微复杂一些的计算器，特别是那些称为“财务计算器”的，只要按一个键就能计算趋势直线。直接计算增长率，要求计算器能够计算一个数的 n 次方根，如第六章(2)增长率例子中的

$$\sqrt[80]{\frac{X_n}{X_i}}$$

虽然任何能够计算对数的简单的计算器(大多都能)都可以按该节所描述的方式加以使用。任何能够计算趋势直线的计算器也可以计算 n 次方根，因而使计算第六章表6.6所描述的对数一线性趋势很为方便。第七章叙述的主要方法，即简单的线性回归和相关分析，可以从许多为科学和统计分析之用而出售的计算器中求得，而第八章所描述的对样本资料的估计方法，可以用与卡方和正态分布的表格有关的计算器计算。只有当人们想要甚至不用这些表格进行计算时，才有必要去购买更高级的计算器，这些计算器能够计算许多由于太复杂而没有在本书讲到的统计分析。

因此历史学家似乎只需决定他要做什么，就可以看出他

---

① 指那种只有加、减、乘、除、累加等12种简单功能的计算器。——译者

需要购买多么复杂的计算器。可是未必尽然，因为小型计算器有一个对历史资料的分析很重要的缺点（假如历史学家主要是想用计算器来帮助阅读，检查和计算书本和论文所提供的资料，或许问题不那么严重）。问题在于许多，如果不是全部，历史问题只能通过对相当大的资料集以及对许多各自拥有若干变量的个案的分析才能解决。举个最简单的例子，似乎有必要计算表3.1所展示的有关1086年埃塞克斯郡牧猪的平均数和标准差。就许多计算器而言，这类计算可以只要通过输入里特尔庄园的牧猪数1200，再按下标有Σ_x的键，并对表中的每一个数字重复这些运算。这样标有$\overline{X}$和S的键就将给出结果。这看起来似乎简单，但值得指出的是，方才描述的这些运算需按键178次（每个数字一次），Σ_x 49次，而$\overline{X}$和S各1次。在这么多次按键中，任何一个错误必须立即注意到，以便改正；否则，答案将是错误的，但是检验结果的唯一办法是重新按键178次。像计算牧猪数目与埃塞克斯郡庄园的某些其他特征之间的相关关系这类更复杂的运算，可能会要求按键455次，虽则确切的按键次数将取决于计算器的设计。

如上述这个简单的例子所示，大多数计算器的主要缺点是输入资料的过程繁琐，而且没有所输入资料的永久性记录可资检查是否已造成错误。在许多资料规模较小的数学例子中这还无关紧要，但对于像历史这样与"资料有密切关系"的课题，而非"咀嚼数字"，亦即对许多小量资料进行复杂计算时，它就成为一个严重的问题了。这个问题可以克服，但克服的代价昂贵；虽然附着于计算器的打印机可以买到，但通常它们的价格是计算器本身的几倍，并使用特殊而又昂贵的纸张，这样就降低了小型计算器所拥有的便于携带和容易使用的优势。克

服这一缺点还有其他方法，但由于需将同一资料分析若干次而颇为麻烦。一些高级的计算器允许资料像“程序”一样(若干组处理资料的指示)被记录在磁带上以备后用。但是这些方法没有一个价廉，没有一个便于使用。

9.2 设备的选择：计算机

由于这些原因，历史学家在分析任何数量的资料时应考虑采用计算机是否明智。这取决于使用计算机是否容易，取决于资料是否需要储存，取决于是否需要重复或复杂的计算。由于使用计算机的情况及费用在不同的教育机构、不同地区和不同国家内的差别甚大，对使用计算机这个问题无法一概而论，所可说的是使用计算机已变得日益容易和价廉了。

为了懂得为什么储存资料和计算性质是重要的，有必要对组成一台电子计算机的部件略知一二。一台计算机主要由3部分构成：执行算术运算的中央处理机；当中央处理机执行指示去处理材料并产生结果时，存储资料、结果和指示集(程序)的存储器；以及将资料传送入存储器并打印出结果的输入—输出设备。历史学家只需知道中央处理机的存在就行了；只有当他打算应用一台微型计算机或小型计算机，或用复杂的分析程序去处理一个巨大的资料集时，存储器才会对他有影响。然而任何使用计算机的人都会受到输入—输出设备及所能应用的程序类型的影响；后者将在下面(4)节里讨论。

输入设备的效用是将资料变成一种可以被计算机处理的形式，计算机把信息作为一系列电荷存储起来。输出设备所做

的恰好相反，使我们能知道处理的结果。输入计算机最常见的方法是通过把它们打成穿孔卡片（如图 9.1 所示）使资料和程序成为“机器可读”的形式。信息通过一些小矩形孔或合并的孔表述在这类卡片上，这些矩形孔被一台附属于计算机的读卡机阅读并翻译成数目、字母或标点符号。在图9.1中，所记录的信息是表4.1中所示的第一艘商船的详细情况。记录的第一个数是1697，这个数字的第一位数1由第1列第1行所穿的孔表示，第二位数6由第6行第2列的穿孔表示，第三位数在第9行第3列，第四位数在第7行第4列。所以，卡的每一列代表一个字符，而字符按穿孔所在的行来区分。每艘船有一张单独的卡片，而且每一个变量在每一张卡片相同的列上被穿孔。由于在卡上只印有10行，所以我们只能表示10个字符。为了克服这一限制，可以用同一列中的两个或更多的穿孔表示某些字符。例如，名称“FLOUD 001”在卡的右边被穿孔，作为一个标识。字母F由第6行的一个穿孔以及卡最上边的一行即通常被称为“+穿孔”位置上的一个穿孔所表示。字母L由第3行中的一个穿孔以及“-穿孔”位置即“+穿孔行”和“O”行之间一行中的一个穿孔表示。通过使用+的和-的穿孔行，连同卡上的其他行，所有字母，所有数字和许多标点符号都可以由独特的合并穿孔表示。因而可以把字母和数字的字符穿孔而不致产生任何混乱。在卡上被穿孔的同时，卡最上方的字符也由卡片穿孔机打出，这样就使得操作员可以检查他所穿的字符是否正确。

穿孔卡片的应用便于许多目的，它们很易被改正和复制，而且所包含的信息在卡片上一目了然。然而当所需卡数较大，如记录大量资料时，这些优点则为卡片本身的分量和易

1697 HOME NG 44 3 FLOUD 001

IBM

图9.1 穿孔卡片

遭损坏所超过。由于这个原因，长期以来穿孔卡通常只用于资料处理的初级阶段；它们只由计算机阅读一次，然后计算机将信息存储在另一种存储介质上——通常是磁带或磁盘。如同磁带录音机中的磁带那样，信息作为电荷被存储在磁带和磁盘上。很自然的一种发展是将资料直接转换到磁带或磁盘上，通过在一台与计算机相连的、很像电动打字机的设备——终端机上打字就可以做到这一点。进一步的发展是用终端机将资料记录在磁带盒上，终端机的形状和重量与电子打字机相似，但它可以将所有输入的，打在纸上或磁带上的东西记录在磁带盒上。这类终端机在各处都可方便地使用，而且日后磁带盒可以拿来为计算机阅读。

计算硬件的这些发展对历史学家来说具有极为重要的意义。这是因为应用于历史学的研究方法和统计方法，与历史资料的收集和分析有密切的关系；使这类工作变得更容易和更省钱的任何事物，对历史学家都是重要的。直到不久以前，历史学家还需要去一个存有历史记录的机构或其他能提供这类记录的地方，吃力地复印出他所想要使用的资料，再将这一复本交给卡片穿孔机的操作员，然后把穿孔卡片喂入计算机。现在，可以把终端机送到存有历史记录的机构里，把所有资料都打在磁带盒或小型磁盘上，以后再输入计算机里，如有必要，甚至可以经由一条电话线传送到计算机上。

同样，输出硬件设备上的发展也对历史学家大有帮助。他不仅可以得到打印出来的结果，而且现在还可以运用计算机绘制的地图和图形来显示他的结果，并且在缩微胶卷和缩微胶片上永久保存这一结果。他可以在一个视频显示装置（这一装置简称VDU，一般由一台电视机荧光屏和一个与之相连的

键盘组成)上显示资料，并可以通过键盘修改资料或向计算机发出指示。他还可以从资料中抽选样本，将它们储存在磁带或磁盘上以备将来分析之用，同时计算机复制多份副本的能力使他的资料免遭意外损害或破坏。

各种输入—输出设备因而种类很多，而新的设备又层出不穷。由于计算硬件的进一步发展，它们的使用也变得更为容易。直到不久以前，计算机与输入—输出设备都紧密置于一处，通常放在大学或学院的计算机中心里。穿孔卡片要送到计算机中心，或在那里制作；然后再将它们喂入计算机；其结果就近被打印出来，常经过长时间耽搁再由研究人员收集起来。因此，计算机的使用者，差不多在一切场合，都受到就在他身边的计算机的限制。他不能便利地应用适合于他的研究目的，但又不是为他必须使用的特定的计算机系统设计的计算机程序，而且将穿孔卡片或甚至磁带传送到远方的计算机里非常麻烦。

由于被称为"计算机网络"的发展，许多上述的困难已经消失，或将在今后几年里消失。在一个特定的机构里，输入—输出设备在装置上与计算机相互分离的情况目前已十分普遍，这些输入—输出设备靠近使用者并经由一条电话线与计算机相联系。在大学和学院之外，现在许多学校经由电话线与一台中心计算机相联系；它们可以传送或接收资料和结果，并能在中心计算机里的磁带和磁盘上存储信息。最后，电话线还将一个特定国家之内的或两个国家之间的计算机本身联系起来。这使得计算可以在最有效地做这类工作的地方进行，这些地方常常远离使用者。例如，专门绘制图表或制作缩微胶片的输入—输出设备可以安置在少数几个专门的计算

机里，并有时为分散的使用者使用。因为电子传递的速度极快，坐在终端机前的使用者错认为他在单独地使用着计算机，而事实上几十或几百个使用者可能正在自己的办公室、学校、车间，甚至在家里同样在使用。

9.3 为计算机准备历史资料

直到不久以前，许多人仍然认为计算机最大的好处就在于它处理数字的能力。这并不是因为计算机还不能存储和处理文字和字母形式的材料，而是因为计算机在早期的明显应用主要在数字处理的领域，因此，在设计计算机、它们的输入—输出设备，以及向计算机发出指示集的程序语言时，大部努力都投入提高数字处理的速度和效率上去了。虽然计算机一直能够处理字母形式的而不仅是数字形式的材料，这项工作却常常十分难办，并在计算中要求比大多数历史学家所想要具备的更多的专门知识。

然而，在历史研究中应用计算机的潜力与上述这些缺点同样明显，因此许多历史学家和其他社会科学家努力使用计算机。使文字形式的历史材料可用于为处理数字而设计的种种方法的最主要的途径，是通过称为"编码"的过程把文字转换成数字。例如，让我们假设历史学家想要研究18—19世纪军队召募新兵的情况，并特别对应募者的职业和出生地感兴趣。在原始的征兵记录中，详细情况按若干种方式记录，表9.1显示的是一个英国征兵记录的例子。为了将这类资料转换成数字形式，每一个职业和出生地都必须被指定一个数字，这个数字将被记录在纸上和穿孔卡片上以输入计算机；这些

一般编码表

姓名	0 1 2 数字 O I Z 文字	日期 29/1/1979
地址	程序名称	

1–10	11–20	21–30	31–40	41–50	51–60	61–70	71–80
W05412̄72	17 71 17̄66	28 3	5 7 50	24			
W0541272	19 11 1766	27 2	5 8 50	4			

图9.2　准备被穿孔的表9.1中一部分的编码表格

W054/272 17 11 1766 26 3 5 7.50 24

资料

W054/272 19 11 1766 27 2 5 .50 41

资料

IBM UNITED KINGDOM LIMITED 886-X21009 UK UNIVERSITY OF LONDON

图9.3 表9.1中部分资料的编码形式,按固定格式穿孔

表9.1　18世纪60年代英国军队的征兵记录：1755—1831年第三炮兵营记录簿

征兵日期	姓　名	年龄		身高		职　业	出生地
		年	月	英尺	英寸		
1766.11.17	B.布思	26	3	5	7.5	锯木工	芝彻斯特
1766.11.19	G.惠丁	27	2	5	8.5	打马蹄铁工	北安普敦
1766.11.26	E.布莱克	22	9	5	6.5	面包师	特威克斯布里

连续的步骤如图 9.2 和图 9.3 所示。

然而，马上就产生了如何指定这些数字的问题。一种极端的方法是为每一个职业和出生地给定一个独特的数字，只有再一次碰到这一职业或出生地时这个数字才会重新出现。例如，所有修车人可编码为7，而在埃塞克斯郡的芝宾盎格一地总以215出现。然而，以这种方式指定数字是极为费力的，因为可能出现的职业和出生地太多，每一项都可能出现在记录中因而需要被指定一个数字；甚至在1841年英国人口普查中有案可查的职业就超过900个，而可能的出生地的数目还要大得多。要记住这样一串名称和数字是不可能的，因此像图9.2那样记录资料的方法既费时又容易出错。况且其结果将是一长列数字而不是一长列职业，对材料的历史分析将无法进行。

为了克服这些困难，许多准备将历史材料转换成数字形式的编码方案采取了把材料的逻辑分类体现在编码过程中的进一步的办法。这样做的另一个好处是可以把编码方案的逻辑用于对材料的分析。表9.2显示出一些可能的编码方案。方案A是只按照在资料中发现的顺序为各种职业指定数字的结果，而方案B则是按字母的顺序排列职业。进一步的逻辑分

表9.2 对个人职业可能的编码方案

1	铸工	1	制筐工	01	制椅工	01	铜匠	1	铜匠
2	铜匠	2	装订工	02	锯木工	02	铸工	1	铸工
3	锯木工	3	制椅人	03	船木匠	03	锡匠	1	锡匠
4	船木匠	4	成衣工	04	修车工	04	船木匠	1	船木匠
5	漆工	5	铜匠	11	铜匠	05	装配工	1	装配工
6	枪炮工人	6	鞋匠	12	枪炮工人	21	漆工	2	漆工
7	修车工	7	钉马蹄铁工	13	铸工	22	管子工	2	管道工
8	鞋匠	8	手套工	14	锡匠	23	石工	2	石工
9	装订工	9	枪炮工人	21	成衣工	24	锯木工	2	锯木工
10	手套工	10	理发师	22	手套工	31	制椅工	3	制椅工
11	制筐工	11	铸工	31	制筐人	32	枪炮工人	3	枪炮工人
12	成衣工	12	驳船船员	32	装订工	33	成衣工	3	成衣工
13	驳船船员	13	石工	33	鞋匠	34	手套工	3	手套工
14	锡匠	14	漆工	34	漆工	35	制筐工	3	制筐工
15	管道工	15	管道工	35	管道工	36	装订工	3	装订工
16	制椅工	16	装配工	36	装配工	37	鞋匠	3	鞋匠
17	钉马蹄铁工	17	锯木工	37	石工	38	修车工	3	修车工
18	石工	18	船木匠	41	钉马蹄铁工	41	钉马蹄铁工	4	钉马蹄铁工
19	装配工	19	修车工	42	理发师	42	理发师	4	理发师
20	理发师	20	锡匠	43	驳船船员	43	驳船船员	4	驳船船员
	A		B		C		D		E

类被用于方案C和D；首先，方案C根据涉及不同种类的原料来区别不同的职业。这样，所有与木材有关的职业被给予01—09之间的数字，所有涉及服装的职业被给予21—29之间的数字，如此等等。从41—49的数字是余下的种类，这些职业不直接涉及原料的加工。对比之下，方案D则按照它们所属的行业对职业进行分类。这样金属和造船业得到01—09之间的数字，建筑业的数字为21—29，而其他制造行业被给予31—39之间的数字。以同样的型式还可以想出其他的逻辑方案。图9.2和图9.3所示为按方案D编码的资料。

进一步的方法是像方案C和D那样将职业分成组，但是对于落入每一组中的所有职业只指定一个独特的数字；例如，所有建筑行业的职业的号码都是2，所有列入其他制造行业的职业号码都是3，等等。正如方案E所示，这样一种按类分组易于记忆，便于编码并且不大可能发生转换的错误。

因此，编制合适的编码方案须用很大的智巧；正确设计的编码方案使历史学家受益非浅。① 但是，如我们将在下面所看到的那样，如果在把历史记录转换成机器可读的计算机文件的过程中过早地进行编码，这亦能是有害的。

编码可按两种方式之一进行。第一种是按手工进行，即当历史学家在阅读他的原始记录并将此记录抄录在如图9.2的通常被称为“编码表格”的纸上，一个穿孔机操作员可以据此工作。这样做的一个好处是，由于数字编码一般要比原始

---

① 参见E. A. 里格利编：《十九世纪的社会》中W.A. 阿姆斯特朗“对有关职业的信息的应用”一章(W. A. Armstrong's Chapter 'The use of information about occupations' in E. A. Wrigley (ed), Nineteenth-Century Society)，剑桥：剑桥大学出版社，1972年。

信息所占的空间少，抄录和穿孔的量也要比如果记录以字母形式抄录小。这样做可能使转换过程花费较少，对于所有，例如，与1名士兵有关的信息必须记录在1张穿孔卡片上这种情况也易于处理。

然而，这种编码方法也伴有相当严重的困难和代价，通常称为“预先编码”，因为它在资料成为一种机器可读形式之前已经编制了。在比表9.2所示要长得多的一览表中寻找合适的数字编码所产生的困难可能要超过转换费用在表面上的节省，犯错误的可能性也很大。然而最重要的是，在转换过程中编码方案和所指定的编码数都是固定的，一旦资料按一特定方案编码以后，要使之分解并恢复原状，或者在另一种合乎逻辑的基础上重新整理资料使之适合一种新的分析方法，都是非常麻烦的。在极端而论，像方案E这样的分组编码方案要做到这一点是不可能的，因为原始材料中的细节已经不可挽回地丢失了。

由于上述原因，历史学家正愈来愈多地，并且明智地利用第二种编码方法。现在计算既适宜于数字，也适宜于文字，并且输入文字形式的信息像输入数字形式的信息一样容易。这使历史学家没有必要在穿孔之前对他的资料进行编码；他只需以文字形式（如果有必要，完全按照原始历史文件的形式）抄录资料并加以穿孔。如图9.4所示，这里资料被穿孔并在每一项信息之间增加了一个/号，它使计算机程序可以区分有关身高和有关职业的信息。一旦资料成为一种机器可读的形式，那么编写一个能阅读资料，并将其按所要求的任何合乎逻辑的编码方案进行编码的计算机程序并不困难；况且，由于原始信息被保存起来了，这种由计算机完成的编码可以毫不费力

地重复无数次。如果信息被打印出来或者显示在一台终端机上，它们很容易被理解，各种错误也会很快被发现；然后使用在大多数计算系统上都具有的校订命令对此加以改正。

资料成为机器可读形式以后再对其进行编码的另一项好处，通过比较图9.3和图9.4就可以看出来。在图9.3中，某一种类的信息，例如职业信息，在每一张卡片的相同位置被穿孔，所编制的计算机程序将所有处于这一位置的编码当作职业编码来译释；这样做浪费空间，因为卡上一定数量的列必须为可能并不使用的编码留出之，这样就容易将编码置于错误的列上。这种被称为“固定格式”输入在图 9.4 中被没有上述缺点的“自由格式”输入所代替；如果以后资料需要成为一种固定的格式，可专门编制程序将它们转换成这种形式。

因此，历史学家能够通过谨慎地使用自由格式输入和像图 9.4 中/这类的符号，在将材料输入计算机的整个过程中保持原有历史文件的特征和形状。当输入完成后，就可以进行编码，并被视为历史分析中的一个明确的步骤。

9.4 运用计算机分析历史资料

资料一旦被转换或机器可读的形式而且如有必要被编码之后，分析就可以开始了，从本书前面几章所描述的重整、分类和描述开始，并通过本书后面几章所讨论的和未涉及的较复杂的分析方法继续方析。所有这些运算工作，无论简单或复杂都必须作为由历史学家向计算机发出的系列指示的结果来进行：“计算机分析”这个术语仅是“使用计算机进行分析”的简称，选择使用哪种分析方法仍然完全由历史学家来决定。

W054/272/17/11/1766/BRIAN/BOOTH/26/3/5/7.5/SAWYER/CHICHESTER

资料

W054/272/19/11/1766/GEORGE/WHITING/27/2/5/8.5/FARRIER/NORTHAMPTON

资料

IBM UNITED KINGDOM LIMITED　　　　1BBG X21009 UK　　　　UNIVERSITY OF LONDON

图 9.4　图9.2的非编码形式，按自由格式穿孔

计算机所能遵循的系列指示就是“程序”。编出计算机可以立即执行的程序是可能的，但这是一项极为枯燥费事的工作，因为程序规定了所需采取的每一步的大量详尽的细节。这正像我们不去告诉某人穿过街道，而是向他发出要达到这一目的所需的每一个肌肉活动的详细指示。由于所要求的这些活动的数量十分巨大，并多次重复，那么很清楚只要能讲“穿过街道”并知道这句话之后将会有一系列活动跟上去就可以了。与此相同，能编出一个其中包含找出一个数字的向量之和的程序就可以了，而毋需确切地指定怎样做到这一点的指示。为达到这一目的，程序用所谓“高级语言”编写，然后它们由一种“编译程序”翻译成可由计算机一步步执行的非常详细的系列指示。这种语言中最著名的是主要为科学和统计而设计的**FORTRAN**① 和**ALGOL**②，以及为商业用的**COBOL**。③ 然而还有许多这类高级语言，其中不少专用于某种特定的计算机，或者甚至特定的计算机中心；因此历史学家必须寻求关于既可得到而又最能满足他的需要的语言的建议。

以大多数高级语言都可编写能够运算本书所讲过的任何描述性或分析性统计方法的程序。然而，历史学家打算这样做无疑是徒费精力。为制作一个列联表或得出一组回归结果，即使运用高级语言写出的程序也既复杂又冗长；但一个程序（或许加以细微的修改）便能一再应用于不同的资料集。因此，被称为“程序包”的各种程序集为人们编出用以完成除了最专

① FORTRAN (formula translation)，公式翻译程序语言。——译者

② ALGOL (algorithmic language)，算法语言。——译者

③ COBOL(common business oriented language)，面向商业的通用语言。——译者

门的分析方法之外的所有分析方法。使用一个程序包的历史学家只需确定他所想要分析的资料的真正性质，然后发出“列联表”或“根据 X 求 Y 的回归”等指示。这些指示将由程序包翻译成以高级语言写成的各种程序，再由编译程序翻译成由计算机执行的各种机器指示。

例如，第七章所描述的表4.1中商船资料的相关和回归分析，就可以通过使用称为SPSS（社会科学用统计程序包）的程序包来完成。它是应用得最广的统计程序包，大多数大学的计算机上都有这种程序包可用。还有许多其它类似的程序包，一些是通用性的，另一些则是为特定类型的分析设计的，如TSP（时间数列程序包）或者COCOA（语词索引和文本处理程序包）；再重申一遍，有关恰当使用程序包的问题可以向进行分析工作的计算机中心咨询。

一个用于SPSS的系列指示包括4部分。第一部分称为JCL（作业控制语言），它并不是SPSS的一部分，只是告知所使用的特定计算机将使用SPSS；具体指示因计算机而异。其它3部分对任何SPSS的“作业”（即调入计算机的一组程序和资料）都是常用的。其一为控制资料输入和描述资料的系列指示，其二为调用特定类型的统计分析的系列指示，其三为需要分析的资料集。整个指令集如图9.5所示，每一行指令都有一条注释，而作业的计算结果，包括SPSS的控制卡片和相关与回归分析的结果见本章末尾的附录。

SPSS和类似程序包使得历史学家无需自己编写统计分析程序；这不仅因为自己编写程序非常费时，而且因为在效率和甚至准确性方面非专业人员编写的程序肯定要比若干组编写过程序的专业程序编制员所得出的结果差得多。历史学家

需要学会怎样以控制卡片和资料的形式向程序包发出指示，当然他还需具有足够的统计学知识，以便从程序包中选择适当的统计例行程序。然而所有这些工作并不要求历史学家具有编写计算机程序的能力。

然而，历史学家有一个需要获得这方面的专门知识的原因。图9.5和附录所示的例子所使用的是一个非常简单明了的资料集，在开始对它分析之前并不需要进行任何处理。但当需要使用较复杂的资料集，尤其是文字需按上一节所述方式进行处理，编码和分析时，在重整资料调入 SPSS这类程序包进行分析之前就需编写一些用于特殊目的的程序来完成这些工作。许多程序包，包括 **SPSS**，都能接受比图 9.5 和附录所用的复杂得多的资料集，但是目前很少数程序包能处理将教区记录簿、人口普查统计表或征兵记录这类历史文件直接转为机器可读形式时所产生的复杂情况。因此，历史学家需要获得足够的编写程序能力，自己来完成这项工作，或至少充分地了解这些问题以便向专业程序编制员作出恰当的指示。然而，由于不仅照管分析而且还照管在分析之前储存和处理资料的新型程序包的发展，现在这种需要正在减小。这类程序包被称为"资料库管理系统"，已经用于历史研究中，也用于许多像编制索引、目录分类和参考书目这些相关领域。与统计程序包相结合，它们为历史研究提供了强有力的帮助。任何打算记录和分析大量资料的历史学家，应考虑应用这类系统，并征求有关专家关于它们的意见。

然而，上段中的评述仅适用于历史学家使用计算机的一小部分情况（虽然是正在扩大的一小部分）。在大多数项目中，资料集的规模和复杂性不至于产生巨大的困难，而像 SPSS

语句	行号	说明
作业 RCF2 2534 SPSS 运行，存盘	——0001	作业控制语言
程序 控制上限，时间10，存储限制 200K	——0002	
打印机 3K	——0003	
技术报告 隔夜	——0004	
遭破坏可恢复	——0005	
SPSS(SYSIN %H+)	——0006	
运行名称 船员人数对船舶吨位的回归分析	——0007	作业题目
变量表 船舶吨位 船员人数	——0008	变量表＝本例中2个
输入方法 卡片	——0009	输入方法——可以是卡片、磁带、磁盘
输入格式 固定(F5.0,F3.0)	——0010	在每一卡片上数据的格式和位置
个案数目 25	——0011	个案数目
回归 变量＝船舶吨位，船员人数/	——0012	求回归指示
回归＝船员人数对船舶吨位 (2)残差＝0/	——0013	
统计 全部	——0014	计算若干可选择的统计指示
读入输入数据	——0015	读入以下数据指示
44 3	——0016	数据
144 6	——0017	
150 5	——0018	
236 8	——0019	
739 16	——0020	
970 15	——0021	
2371 23	——0022	
309 5	——0023	
679 13	——0024	
26 4	——0025	
1272 19	——0026	
3246 33	——0027	
1904 19	——0028	
357 10	——0029	
1080 16	——0030	
1027 22	——0031	
45 2	——0032	
62 3	——0033	
68 2	——0034	
2507 22	——0035	
138 2	——0036	
502 18	——0027	
1501 21	——0038	
2750 24	——0039	
192 9	——0040	
完成	——0041	作业结束指示
+	——0042	作业控制语言
11 *	——0043	

图 9.5 一份 SPSS 作业输入

资料据表 4.1。

这类程序包中的资料管理能力，助以简单的预备程序，将能满足大多数需要。然而重要的是，历史学家应认识到潜在的问题，并应在记录他的资料之前解决这些问题。否则，当一种稍微不同的记录方法可以会使资料很容易符合分析程序包的种种要求时，为了进行分析可能需要花费大量的时间和精力去重新整理资料。所有这些程序包在已出版的手册中都有完整的文字说明，这些手册在提供使用这些程序包的任何计算机中心里都可以得到，应该早期阅读它们。

计算为历史学家提供了巨大的机会，因为它使历史学家能够按照以前根本不可能的方式去组织，分析和理解历史资料。当然，不能保证这种研究的结果将是重要的或有价值的，但是很清楚地需要的是，新的资料范围应对历史研究开放。至于历史学家怎样善于运用新的资料以及他们怎样善于运用本书所叙述的所有统计方法，只有等做了工作以后才能予以评价。

图 9.5 中 SPSS 作业的计算输出

社会科学用统计程序包　　01/31/79

OS/360系统适用的 SPSS版本 H.　7.2号　1977.12.5

缺省的空间分配　允许使用　64种转换

工作空间　44800字节

转换用空间　6400字节　256个重编码值＝滞后变量

1024个判别或运算操作

运行名　船员人数对船舶吨位的回归分析

变量表　船舶吨位　船员人数

输入介质　卡片

输入格式　固定(F5.0,F3.0)

按照你的输入格式,变量须以下面方式读入:

变量	格式	记录	列
船舶吨位	F5.0	1	1—5
船员人数	F3.0	1	6—8

输入格式含2个变量,将读入2个值。

输入格式为每个个案提供一个记录(卡片),一个记录最多可使用8列

个案数目　25

回归　变量＝船舶吨位,船员人数/

回归＝船员人数对船舶吨位(2)残差＝0/

统计　全部

*****回归问题需要182字节的工作空间,不包括残差*****

附录

读输入数据

船员人数对船舶吨位的回归分析　　01/31/79

文件　未取名　(建立日期:01/31/79)

变量	平均值	标准差	个案数
船舶吨位	892.7600	965.8430	25
船员人数	12.8000	8.7464	25

船员人数对船舶吨位的回归分析　　01/31/79

文件　未取名　(建立日期:01/31/79)

相关系数

如果某个系数无法计算，

则打印一个99.0000表示。

	船舶吨位	船员人数
船舶吨位	1.00000	0.90936
船员人数	0.90936	1.00000

船员人数对船舶吨位的回归分析　　01/31/79

文件　未取名　(建立日期:01/31/79)

…………多重回归…………　变量表1

回归表1

因变量　船员人数

在步骤1输入的变量值　船舶吨位

多重R　0.90936

R平方	0.82693	方差分析	DF	平方和	平均平方值	F
调整的R平方	0.81940	回归	1.	1518.24230	1518.24230	109.98371
标准误差	3.71693	残差	23.	317.75770	13.81555	

——方程中的变量—— ——不在方程中的变量——

变量	B		BETA	标准误差	B	F	变量 BETA 在偏相关容差'F
船舶吨位	08234901	D-02	BETA	标准误差	B	F	
(常数)	5.448210		0.90936	0.00079		109.894	

已达到的最大步骤数

无法计算的统计值都以打印全'9'表示

船员人数对船舶吨位的回归分析 01/31/79

文件 未取名 (建立日期=01/31/79)

……………………多重回归…………………… 变量表1

回归表1

因变量 船员人数

一览表

变量	多重R	R平方	R平方变化	原始R	B	BETA
船舶吨位	0.0936	0.82693	0.82693	0.90936	0.8234901D02	0.90936
(常数)					5.448210	

船员人数对船舶吨位的回归分析 01/31/79

……………………回归问题要求2232字节的工作空间，包括残差……………………

船员人数对船舶吨位的回归分析 01/31/79

文件　　未取名　　（建立日期＝01/31/79）

……………………多重回归……………………

因变量：　船员人数　　从　　变量表1

　　　　　　　　　　　　　　回归表1

序号	船员人数的观察值	船员人数的预测值	残　差	画出标准残差曲线 −2.0　−1.0　0.0　1.0　2.0
1	3.000000	5.810546	−2.810545	
2	6.000000	6.634035	−0.6340357	
3	5.000000	6.683445	−1.683445	
4	8.000000	7.391646	0.6083534	
5	16.00000	11.53380	4.466198	
6	15.00000	13.43606	1.563936	
7	23.00000	24.97314	−1.973159	
8	5.000000	7.992794	−2.992794	
9	13.00000	11.03971	1.960292	
10	4.000000	5.662317	−1.662317	
11	19.00000	15.92300	3.076996	
12	33.00000	32.17868	0.8213021	
13	19.00000	21.12746	−2.127460	
14	10.00000	8.388069	1.611930	
15	16.00000	14.34190	1.658098	
16	22.00000	13.90545	8.094546	
17	2.000000	5.818781	−3.818780	
18	3.000000	5.958774	−2.958774	
19	2.000000	6.008183	−4.008183	
20	22.00000	26.09309	−4.093105	
21	2.000000	6.584628	−4.584626	
22	18.00000	9.582130	8.417870	
23	21.00000	17.80879	3.191203	
24	24.00000	28.09418	−4.094187	
25	9.000000	7.029311	1.970689	

通过比较个案次序（按序号）的德宾-沃森检验残差

变量表 1　　回归表 1　　德宾-沃森检验　　1.91190

船员人数对船舶吨位的回归分析　　01/31/79

文件　　未取名　　(建立日期 = 01/31/79)

…………画出标准残差曲线(如下)…………

预测的标准化因变量(交叉)…………

因变量：　船员人数　　变量表 1

回归表 1

行、列Y：外值(-3.0, 3.0)　　行、列X：内值(-3.0, -2.05)或(2.95, 3.0)

```
船员人数对船舶吨位的回归分析          01/31/79
所需主机时间      0.38 秒
                    完成
        作业正常完成          已处理 10 张控制卡片
                              发现错误 0 个
           本次运行需要0.1磁道的暂存空间
```

参考书目

由于运用本书所描述的各种方法出版的著作和文章已很多，这里不可能一一列举，甚至列举其中的大部分也不可能。任何参考书目都必须是有选择性的。第一部分列举了一些其论题也包括在本书中的著作，虽然它们的侧重点和所举例子不同。其中不少著作对概率和抽样理论以及较高级的统计方法的叙述比本书所能作出的要多得多。第二部分包括讨论计量方法对历史学问题的应用，以及给出在不同的历史研究领域和历史时期应用计量方法例子的著作。几乎所有这些著作都包含更广泛的参考书目，因此建议本书读者从第二部分列举的著作中寻找合乎自己兴趣的其它书籍和文章。

1．有关统计学和计量方法的导论性书籍

R. G. D. 艾伦：《经济学家使用的统计学》(Allen, R. G. D., *Statistics for Economists*)，伦敦：Hutchinson University Library, 1966 年。

H. M.布拉洛克：《社会统计学》(Blalock, H. M., *Social Statistics*)，纽约：McGraw-Hill 1960 年。

E. 考尔科特：《显著性检验》（Caulcott, E., *Significance Tests*），伦敦：Routledge & Kegan Paul，1973 年。

C. M. 多拉尔和 R. J. 詹森：《历史学家统计学导论：计量分析和历史学研究》（Dollar, C. M., and Jensen, R. J., *Historian's Guide to Statistics: Quantitative Analysis and Historical Research*），纽约：Holt, Rinehart and Winston, 1971 年。

M.德雷克：《历史资料和社会科学，第一卷：历史资料的计量分析》（Drake, M., *Historical Data and the Social Sciences: Vol. 1, The Quantitative Analysis of Historical Data*），米尔顿凯恩斯：Open University Press，1974 年。

J.高尔滕：《社会研究的理论和方法》（Galtung, J., *Theory and Methods of Social Research*），伦敦：George Allen and Unwin，1967 年。

M. J. 莫罗尼：《从数字得到事实》（Moroney, M. J., *Facts from Figures*），哈蒙思沃斯：Penguin Books，1960年。

N. H.尼等人：《社会科学用程序包》（Nie N. H. et al., *Statistics Package for the Social Science*），纽约：McGraw-Hill, 1975 年。

S.西格尔：《行为科学用非参数性统计》（Siegel, S., *Nonparametric Statistics for the Behavioural Sciences*），纽约：McGraw-Hill, 1956 年。

K. A.约曼斯：《社会科学家的统计学：卷 I，介绍性统计学；第二卷，应用统计学》（Yeomans, K. A., *Statistics for the Social Scientist: Vol. I, Introductory Statistics; Vol. II, Applied Statistics*），哈蒙兹沃斯：Penguin Books,

1968。

2. 方法论及计量历史研究的文集

R. L. 安德烈诺编:《新经济史:近来有关方法论的论文》(Andreano, R. L. (ed), *The New Economic History: Recent Papers on Methodology*),纽约:John Wiley, 1970 年。

W. O. 艾德洛特:《历史学中的计量化》(Aydelotte, W. O., *Quantification in History*),马萨诸塞:Addison-Wesley,1971年。

W. O. 艾德洛特、A. G. 博格和 R. W. 福格尔编:《历史中计量研究的范围》(Aydelotte, W. O., Bogue, A. G. and Fogel, R. W. (ed), *The Dimensions of Quantitative Research in History*),伦敦:Oxford University Press, 1972年。

W. O. 艾德洛特编:《议会行为史》(Aydelotte, W. O., *A History of Parliamentary Behaviour*),普林斯顿,新泽西:Princeton University Press, 1977 年。

R. F. 伯克霍弗,jun.:《历史分析中的行为方法》(Berkhofer, R. F. jun., *A Behavioral Approach to Historical Analysis*),纽约:Free Press,1969 年。

H. M. 布拉洛克:《社会研究中的方法论》(Blalock, H. M., *Methodology in Social Research*),纽约:McGraw-Hill, 1968年。

A. G. 博格编:《社会和政治史中理论模型的出现》(Bogue, A. G. (ed), *Emerging Theoretical Models in Social and Political History*),贝弗利希尔斯和伦敦:Sage Publica-

tion, 1973 年。

J. 克拉布和 E. K. 朔伊赫编：《历史的社会研究》(Clubb, J. and Scheuch, E. K. (eds), *Historical Social Research*), 斯图加特：Klett-Cotta, 1979 年。

J. 克拉布和 M. W. 特劳高特：《使用计算机》(Clubb, J. and Traugott, M. W., *Using Computers*), 华盛顿特区：American Political Science Association, 1978 年。

L. E. 戴维斯和 D. 诺斯：《体制的变化和美国的经济增长》(Davis, L. E. and North, D., *Institutional Change and American Economic Growth*), 剑桥：Cambridge University Press, 1971 年。

M. 德雷克编：《实用历史学研究》(Drake, M. (ed), *Applied Historical Studies*), 伦敦：Methuen, 1973 年。

M. 德雷克：《历史资料与社会科学：第二卷，历史人口学；第三卷，历史选举学导论；第四卷，历史社会学的实践》(Drake, M., *Historical Data and the Social Sciences*, *Vol. 2, Historical Demography*; *Vol. 3, Introduction to Historical Psephology*; *Vol. 4, Exercises in Historical Sociology*) 米尔顿凯恩斯：Open University Press, 1974 年。

S. L.恩格尔曼和 E.D. 吉诺维斯合编：《西半球的种族和奴隶制度：计量研究》(Engerman, S. L. and Genovese, E. D. (eds), *Race and Slavery in the Western Hemisphere: Quantitative Studies*), 普林斯顿：Princeton University Press, 1975 年。

R. C. 弗拉德编：《计量经济史论文集》(Floud, R. C. (ed.),

Essays in Quantitative Economic History)，牛津：Clarendon Press, 1974 年。

R. W. 福格尔和 S. L. 恩格尔曼合编：《对美国经济史的重新解释》(Fogel, R. W. and Engerman, S. L. (eds.), *The Re-interpretation of American Economic History*)，纽约：Harper and Row, 1971 年。

D. V. 格拉斯和 D. E. C. 埃佛斯利合编：《历史中的人口》(Glass, D. V. and Eversley, D. E. C.(eds), *Population in History*)，伦敦：Edward Arnold, 1965 年。

J. D. 古尔德：《历史中的经济增长》(Gould, J. D., *Economic Growth in History*)，伦敦：Methuen, 1972 年。

T. H. 霍林斯沃思：《历史人口统计学》(Hollingsworth, T. H., *Historical Demography*)，伦敦：Hodder and Stoughton, 1969年。

M. 英泰利盖脱编：《计量经济学的新领域》(Intriligator, M. (ed)., *Frontiers of Quantitative Economics*)，阿姆斯特丹，North Holland，1971 年。

F. 艾西格勒编：《近代前的经济史和社会史中的计量方法》(Irsigler, F.(ed.), *Quantitative Methoden in der Wirtschafts und Sozial Geschichte der Vorneuzeit*)，斯图加特：Klett-Cotta, 1978 年。

G. 柯根和 P. 穆利奥斯合编：《历史中的计量化》(Kurgan, G. and Moureaux, P. (eds)., *La Quantification en Histoire*)，布鲁塞尔，Brussels University Press, 1973 年。

E. 列奥·拉杜里：《历史学家的领域》(Leroy Ladurie, E., *Le*

Territoire de l'Historien)，巴黎：Gallimard, 1973 年。

D. S. 兰德斯和 C. 蒂利：《作为社会科学的历史学》(Landes, D. S. and Tilly, C., *History as Social Science*)，恩格利沃德克利夫斯，新泽西：Prentice Hall, 1971 年。

C. H. 李：《经济史研究的计量方法》(Lee, C. H., *The Quantitative Approach to Economic History*)，伦敦：Martin Robertson, 1977 年。

R. D. 李编：《历史上的人口模式》(Lee, R. D. (ed.), *Population Patterns in the Past*)，纽约：Academic Press, 1977年。

S. M. 利普塞特编：《政治学与社会科学》(Lipset, S. M. (ed.), *Politics and the Social Sciences*)，纽约：Oxford University Press, 1969 年。

S. M. 利普塞特和 R. 霍夫施塔特合编：《社会学与历史学：方法论》(Lipset, S. M. and Hofstadter, R. (eds.), *Sociology and History: Methods*)，纽约：Basic Books, 1968 年。

V. R. 洛温和 J. M. 普赖斯：《历史的维度：历史学中计量研究的材料、问题和机会》(Lorwin, V. R. and Price, J. M., *The Dimensions of the Past. Materials, Problems, and Opportunities for Quantitative Work in History*)，纽黑文和伦敦：Yale University Press, 1972 年。

P. D. 麦克莱兰：《历史中的因果解释和模型的建立，经济学和新经济史》(McClelland, P. D., *Causal Explanation and Model Building in History, Economics and the New Economic History*)，伊撒卡，纽约：Cornell University

Press, 1975 年。

D. N. 麦克洛斯基编，《论成熟的经济：1840 年以后的英国》(McCloskey, D. N.(ed.), *Essays on a Mature Economy: Britain after* 1840)，伦敦：Methuen, 1971 年。

A. 麦克法兰：《重建历史的社区》(MacFarlane, A., *Reconstructing Historical Communities*)，剑桥：Cambridge University Press, 1978 年。

·R. 梅里特和 S. 罗干：《比较国家：在跨国家研究中计量资料的应用》(Merritt, R. and Rokkan, S., *Comparing Nations: The Use of Quantitative Data in Cross-National Research*)，纽黑文，康涅狄格：Yale University Press, 1966年。

P. 奥布赖恩：《铁路的新经济史》(O'Brien, P., *The New Economic History of the Railways*)，伦敦：Croom Helm, 1977年。

D.K. 罗尼和 J.Q.格雷厄姆, jun.合编：《计量史学：历史资料的计量分析选读》(Rowney, D. K. and Graham, J. Q. jun. (eds.), *Quantitative History: Selected Readings in the Quantitative Analysis of Historical Data*)，霍姆伍德，伊利诺伊：Dorsey Press, 1969 年。

L. F. 施努尔编：《新都市史，美国历史学家的计量研究》(Schnore, L. F.(ed.), *The New Urban History. Quantitative Explorations by American Historians*)，普林斯顿，新泽西：Princeton University Press, 1978 年。

J. H. 西尔比，A. G.博格和 W. H.弗拉尼根：《美国选举行为史》(Silbey, J. H., Bogue, A. G. and Flanigan, W. H.,

The History of American Electoral Behavior)，普林斯顿：Princeton University Press，1978 年。

《计量历史学的研究和社会科学的推理方法》(*Studies in Quantitative History and the Logic of the Social Sciences*)，米德尔敦，康涅狄格：Wesleyan University Press，《历史与理论》增刊 9，1969 年。

R. P. 斯威雷加编：《美国史学中的计量化》(Swierenga, R. P. (ed.)，*Quantification in American History*)，纽约：Atheneum，1970 年。

P. 特明编：《新经济史》(Temin, P. (ed.)，*The New Economic History*)，哈蒙兹伍德：Penguin Books，1973 年。

C. 蒂利编：《对变化中的出生率的历史学研究》(Tilly, C. (ed.)，*Historical Studies of Changing Fertility*)，普林斯顿，新泽西：Princeton University Press，1978 年。

K. A. 瓦赫特尔，及 E. A. 哈梅尔，P. 拉斯利特：《历史社会结构的统计研究》(Wachter, K.A. with E. A. Hammel and P. Laslett, *Statistical Studies of Historical Social Structure*)，纽约和伦敦：Academic Press，1978 年。

J. G. 威廉森：《19 世纪末美国的发展，平衡史》(Williamson, J. G., *Late Nineteenth Century American Development. A General Equilibrium History*)，剑桥：Cambridge University Press，1974 年。

E. A. 里格利编：《英国历史人口统计学导论》(Wrigley, E. A.(ed.)，*An Introduction to English Historical Demography*)，伦敦：Weidenfeld and Nicolson，1966 年。

E. A. 里格利编：《19世纪的社会，应用计量方法研究社会资

料论文集》(Wrigley, E. A. (ed.), *Nineteenth Century Society. Essays in the Use of Quantitative Methods for the Study of Social Data*), 剑桥: Cambridge University Press, 1972 年。

E. A. 里格利编:《识别历史中的人》(Wrigley, E. A. (ed.), *Identifying People in the Past*), 伦敦: Edward Arnold, 1973年。

对数表

	0	1	2	3	4	5	6	7	8	9	平均差								
											1	2	3	4	5	6	7	8	9
10	0000	0043	0086	0128	0170	0212	0253	0294	0334	0374	4	8	12	17	21	25	29	33	37
11	0414	0453	0492	0531	0569	0607	0645	0682	0719	0755	4	8	11	15	19	23	26	30	34
12	0792	0828	0864	0899	0934	0969	1004	1038	1072	1106	3	7	10	14	17	21	24	28	31
13	1139	1173	1206	1239	1271	1303	1335	1367	1399	1430	3	6	10	13	16	19	23	26	29
14	1461	1492	1523	1553	1584	1614	1644	1673	1703	1732	3	6	9	12	15	18	21	24	27
15	1761	1790	1818	1847	1875	1903	1931	1959	1987	2014	3	6	8	11	14	17	20	22	25
16	2041	2068	2095	2122	2148	2175	2201	2227	2253	2279	3	5	8	11	13	16	18	21	24
17	2304	2330	2355	2380	2405	2430	2455	2480	2504	2529	2	5	7	10	12	15	17	20	22
18	2553	2577	2601	2625	2648	2672	2695	2718	2742	2765	2	5	7	9	12	14	16	19	21
19	2788	2810	2833	2856	2878	2900	2923	2945	2967	2989	2	4	7	9	11	13	16	18	20
20	3010	3032	3054	3075	3096	3118	3139	3160	3181	3201	2	4	6	8	11	13	15	17	19
21	3222	3243	3263	3284	3304	3324	3345	3365	3385	3404	2	4	6	8	10	12	14	16	18
22	3424	3444	3464	3483	3502	3522	3541	3560	3579	3598	2	4	6	8	10	12	14	15	17
23	3617	3636	3655	3674	3692	3711	3729	3747	3766	3784	2	4	6	7	9	11	13	15	17
24	3802	3820	3838	3856	3874	3892	3909	3927	3945	3962	2	4	5	7	9	11	12	14	16
25	3979	3997	4014	4031	4048	4065	4082	4099	4116	4133	2	3	5	7	9	10	12	14	15
26	4150	4166	4183	4200	4216	4232	4249	4265	4281	4298	2	3	5	7	8	10	11	13	15
27	4314	4330	4346	4362	4378	4393	4409	4425	4440	4456	2	3	5	6	8	9	11	13	14
28	4472	4487	4502	4518	4533	4548	4564	4579	4594	4609	2	3	5	6	8	9	11	12	14
29	4624	4639	4654	4669	4683	4698	4713	4728	4742	4757	1	3	4	6	7	9	10	12	13
30	4771	4786	4800	4814	4829	4843	4857	4871	4886	4900	1	3	4	6	7	9	10	11	13
31	4914	4928	4942	4955	4969	4983	4997	5011	5024	5038	1	3	4	6	7	8	10	11	12
32	5051	5065	5079	5092	5105	5119	5132	5145	5159	5172	1	3	4	5	7	8	9	11	12
33	5185	5198	5211	5224	5237	5250	5263	5276	5289	5302	1	3	4	5	6	8	9	10	12
34	5315	5328	5340	5353	5366	5378	5391	5403	5416	5428	1	3	4	5	6	8	9	10	11
35	5441	5453	5465	5478	5490	5502	5514	5527	5539	5551	1	2	4	5	6	7	9	10	11
36	5563	5575	5587	5599	5611	5623	5635	5647	5658	5670	1	2	4	5	6	7	8	10	11
37	5682	5694	5705	5717	5729	5740	5752	5763	5775	5786	1	2	3	5	6	7	8	9	10
38	5798	5809	5821	5832	5843	5855	5866	5877	5888	5899	1	2	3	5	6	7	8	9	10
39	5911	5922	5933	5944	5955	5966	5977	5988	5999	6010	1	2	3	4	5	7	8	9	10
40	6021	6031	6042	6053	6064	6075	6085	6096	6107	6117	1	2	3	4	5	6	8	9	10
41	6128	6138	6149	6160	6170	6180	6191	6201	6212	6222	1	2	3	4	5	6	7	8	9
42	6232	6243	6253	6263	6274	6284	6294	6304	6314	6325	1	2	3	4	5	6	7	8	9
43	6335	6345	6355	6365	6375	6385	6395	6405	6415	6425	1	2	3	4	5	6	7	8	9
44	6435	6444	6454	6464	6474	6484	6493	6503	6513	6522	1	2	3	4	5	6	7	8	9
45	6532	6542	6551	6561	6571	6580	6590	6599	6609	6618	1	2	3	4	5	6	7	8	9
46	6628	6637	6646	6656	6665	6675	6684	6693	6702	6712	1	2	3	4	5	6	7	7	8
47	6721	6730	6739	6749	6758	6767	6776	6785	6794	6803	1	2	3	4	5	5	6	7	8
48	6812	6821	6830	6839	6848	6857	6866	6875	6884	6893	1	2	3	4	4	5	6	7	8
49	6902	6911	6920	6928	6937	6946	6955	6964	6972	6981	1	2	3	4	4	5	6	7	8
50	6990	6998	7007	7016	7024	7033	7042	7050	7059	7067	1	2	3	3	4	5	6	7	8
51	7076	7084	7093	7101	7110	7118	7126	7135	7143	7152	1	2	3	3	4	5	6	7	8
52	7160	7168	7177	7185	7193	7202	7210	7218	7226	7235	1	2	2	3	4	5	6	7	7
53	7243	7251	7259	7267	7275	7284	7292	7300	7308	7316	1	2	2	3	4	5	6	6	7
54	7324	7332	7340	7348	7356	7364	7372	7380	7388	7396	1	2	2	3	4	5	6	6	7

	0	1	2	3	4	5	6	7	8	9	平均差								
											1	2	3	4	5	6	7	8	9
55	7404	7412	7419	7427	7435	7443	7451	7459	7466	7474	1	2	2	3	4	5	5	6	7
56	7482	7490	7497	7505	7513	7520	7528	7536	7543	7551	1	2	2	3	4	5	5	6	7
57	7559	7566	7574	7582	7589	7597	7604	7612	7619	7627	1	2	2	3	4	5	5	6	7
58	7634	7642	7649	7657	7664	7672	7679	7686	7694	7701	1	1	2	3	4	4	5	6	7
59	7709	7716	7723	7731	7738	7745	7752	7760	7767	7774	1	1	2	3	4	4	5	6	7
60	7782	7789	7796	7803	7810	7818	7825	7832	7839	7846	1	1	2	3	4	4	5	6	6
61	7853	7860	7868	7875	7882	7889	7896	7903	7910	7917	1	1	2	3	4	4	5	6	6
62	7924	7931	7938	7945	7952	7959	7966	7973	7980	7987	1	1	2	3	3	4	5	6	6
63	7993	8000	8007	8014	8021	8028	8035	8041	8048	8055	1	1	2	3	3	4	5	5	6
64	8062	8069	8075	8082	8089	8096	8102	8109	8116	8122	1	1	2	3	3	4	5	5	6
65	8129	8136	8142	8149	8156	8162	8169	8176	8182	8189	1	1	2	3	3	4	5	5	6
66	8195	8202	8209	8215	8222	8228	8235	8241	8248	8254	1	1	2	3	3	4	5	5	6
67	8261	8267	8274	8280	8287	8293	8299	8306	8312	8319	1	1	2	3	3	4	5	5	6
68	8325	8331	8338	8344	8351	8357	8363	8370	8376	8382	1	1	2	3	3	4	4	5	6
69	8388	8395	8401	8407	8414	8420	8426	8432	8439	8445	1	1	2	2	3	4	4	5	6
70	8451	8457	8463	8470	8476	8482	8488	8494	8500	8506	1	1	2	2	3	4	4	5	6
71	8513	8519	8525	8531	8537	8543	8549	8555	8561	8567	1	1	2	2	3	4	4	5	5
72	8573	8579	8585	8591	8597	8603	8609	8615	8621	8627	1	1	2	2	3	4	4	5	5
73	8633	8639	8645	8651	8657	8663	8669	8675	8681	8686	1	1	2	2	3	4	4	5	5
74	8692	8698	8704	8710	8716	8722	8727	8733	8739	8745	1	1	2	2	3	4	4	5	5
75	8751	8756	8762	8768	8774	8779	8785	8791	8797	8802	1	1	2	2	3	3	4	5	5
76	8808	8814	8820	8825	8831	8837	8842	8848	8854	8859	1	1	2	2	3	3	4	5	5
77	8865	8871	8876	8882	8887	8893	8899	8904	8910	8915	1	1	2	2	3	3	4	4	5
78	8921	8927	8932	8938	8943	8949	8954	8960	8965	8971	1	1	2	2	3	3	4	4	5
79	8976	8982	8987	8993	8998	9004	9009	9015	9020	9025	1	1	2	2	3	3	4	4	5
80	9031	9036	9042	9047	9053	9058	9063	9069	9074	9079	1	1	2	2	3	3	4	4	5
81	9085	9090	9096	9101	9106	9112	9117	9122	9128	9133	1	1	2	2	3	3	4	4	5
82	9138	9143	9149	9154	9159	9165	9170	9175	9180	9186	1	1	2	2	3	3	4	4	5
83	9191	9196	9201	9206	9212	9217	9222	9227	9232	9238	1	1	2	2	3	3	4	4	5
84	9243	9248	9253	9258	9263	9269	9274	9279	9284	9289	1	1	2	2	3	3	4	4	5
85	9294	9299	9304	9309	9315	9320	9325	9330	9335	9340	1	1	2	2	3	3	4	4	5
86	9345	9350	9355	9360	9365	9370	9375	9380	9385	9390	1	1	2	2	3	3	4	4	5
87	9395	9400	9405	9410	9415	9420	9425	9430	9435	9440	0	1	1	2	2	3	3	4	4
88	9445	9450	9455	9460	9465	9469	9474	9479	9484	9489	0	1	1	2	2	3	3	4	4
89	9494	9499	9504	9509	9513	9518	9523	9528	9533	9538	0	1	1	2	2	3	3	4	4
90	9542	9547	9552	9557	9562	9566	9571	9576	9581	9586	0	1	1	2	2	3	3	4	4
91	9590	9595	9600	9605	9609	9614	9619	9624	9628	9633	0	1	1	2	2	3	3	4	4
92	9638	9643	9647	9652	9657	9661	9666	9671	9675	9680	0	1	1	2	2	3	3	4	4
93	9685	9689	9694	9699	9703	9708	9713	9717	9722	9727	0	1	1	2	2	3	3	4	4
94	9731	9736	9741	9745	9750	9754	9759	9763	9768	9773	0	1	1	2	2	3	3	4	4
95	9777	9782	9786	9791	9795	9800	9805	9809	9814	9818	0	1	1	2	2	3	3	4	4
96	9823	9827	9832	9836	9841	9845	9850	9854	9859	9863	0	1	1	2	2	3	3	4	4
97	9868	9872	9877	9881	9886	9890	9894	9899	9903	9908	0	1	1	2	2	3	3	4	4
98	9912	9917	9921	9926	9930	9934	9939	9943	9948	9952	0	1	1	2	2	3	3	4	4
99	9956	9961	9965	9969	9974	9978	9983	9987	9991	9996	0	1	1	2	2	3	3	3	4

反对数表

	0	1	2	3	4	5	6	7	8	9	平均差 1	2	3	4	5	6	7	8	9
.00	1000	1002	1005	1007	1009	1012	1014	1016	1019	1021	0	0	1	1	1	1	2	2	2
.01	1023	1026	1028	1030	1033	1035	1038	1040	1042	1045	0	0	1	1	1	1	2	2	2
.02	1047	1050	1052	1054	1057	1059	1062	1064	1067	1069	0	0	1	1	1	1	2	2	2
.03	1072	1074	1076	1079	1081	1084	1086	1089	1091	1094	0	0	1	1	1	1	2	2	2
.04	1096	1099	1102	1104	1107	1109	1112	1114	1117	1119	0	1	1	1	1	2	2	2	2
.05	1122	1125	1127	1130	1132	1135	1138	1140	1143	1146	0	1	1	1	1	2	2	2	2
.06	1148	1151	1153	1156	1159	1161	1164	1167	1169	1172	0	1	1	1	1	2	2	2	2
.07	1175	1178	1180	1183	1186	1189	1191	1194	1197	1199	0	1	1	1	1	2	2	2	2
.08	1202	1205	1208	1211	1213	1216	1219	1222	1225	1227	0	1	1	1	1	2	2	2	3
.09	1230	1233	1236	1239	1242	1245	1247	1250	1253	1256	0	1	1	1	1	2	2	2	3
.10	1259	1262	1265	1268	1271	1274	1276	1279	1282	1285	0	1	1	1	1	2	2	2	3
.11	1288	1291	1294	1297	1300	1303	1306	1309	1312	1315	0	1	1	1	2	2	2	2	3
.12	1318	1321	1324	1327	1330	1334	1337	1340	1343	1346	0	1	1	1	2	2	2	2	3
.13	1349	1352	1355	1358	1361	1365	1368	1371	1374	1377	0	1	1	1	2	2	2	3	3
.14	1380	1384	1387	1390	1393	1396	1400	1403	1406	1409	0	1	1	1	2	2	2	3	3
.15	1413	1416	1419	1422	1426	1429	1432	1435	1439	1442	0	1	1	1	2	2	2	3	3
.16	1445	1449	1452	1455	1459	1462	1466	1469	1472	1476	0	1	1	1	2	2	2	3	3
.17	1479	1483	1486	1489	1493	1496	1500	1503	1507	1510	0	1	1	1	2	2	2	3	3
.18	1514	1517	1521	1524	1528	1531	1535	1538	1542	1545	0	1	1	1	2	2	2	3	3
.19	1549	1552	1556	1560	1563	1567	1570	1574	1578	1581	0	1	1	1	2	2	3	3	3
.20	1585	1589	1592	1596	1600	1603	1607	1611	1614	1618	0	1	1	1	2	2	3	3	3
.21	1622	1626	1629	1633	1637	1641	1644	1648	1652	1656	0	1	1	2	2	2	3	3	3
.22	1660	1663	1667	1671	1675	1679	1683	1687	1690	1694	0	1	1	2	2	2	3	3	3
.23	1698	1702	1706	1710	1714	1718	1722	1726	1730	1734	0	1	1	2	2	2	3	3	4
.24	1738	1742	1746	1750	1754	1758	1762	1766	1770	1774	0	1	1	2	2	2	3	3	4
.25	1778	1782	1786	1791	1795	1799	1803	1807	1811	1816	0	1	1	2	2	2	3	3	4
.26	1820	1824	1828	1832	1837	1841	1845	1849	1854	1858	0	1	1	2	2	3	3	3	4
.27	1862	1866	1871	1875	1879	1884	1888	1892	1897	1901	0	1	1	2	2	3	3	3	4
.28	1905	1910	1914	1919	1923	1928	1932	1936	1941	1945	0	1	1	2	2	3	3	4	4
.29	1950	1954	1959	1963	1968	1972	1977	1982	1986	1991	0	1	1	2	2	3	3	4	4
.30	1995	2000	2004	2009	2014	2018	2023	2028	2032	2037	0	1	1	2	2	3	3	4	4
.31	2042	2046	2051	2056	2061	2065	2070	2075	2080	2084	0	1	1	2	2	3	3	4	4
.32	2089	2094	2099	2104	2109	2113	2118	2123	2128	2133	0	1	1	2	2	3	3	4	4
.33	2138	2143	2148	2153	2158	2163	2168	2173	2178	2183	0	1	1	2	2	3	3	4	4
.34	2188	2193	2198	2203	2208	2213	2218	2223	2228	2234	1	1	2	2	3	3	4	4	5
.35	2239	2244	2249	2254	2259	2265	2270	2275	2280	2286	1	1	2	2	3	3	4	4	5
.36	2291	2296	2301	2307	2312	2317	2323	2328	2333	2339	1	1	2	2	3	3	4	4	5
.37	2344	2350	2355	2360	2366	2371	2377	2382	2388	2393	1	1	2	2	3	3	4	4	5
.38	2399	2404	2410	2415	2421	2427	2432	2438	2443	2449	1	1	2	2	3	3	4	4	5
.39	2455	2460	2466	2472	2477	2483	2489	2495	2500	2506	1	1	2	2	3	3	4	5	5
.40	2512	2518	2523	2529	2535	2541	2547	2553	2559	2564	1	1	2	2	3	4	4	5	5
.41	2570	2576	2582	2588	2594	2600	2606	2612	2618	2624	1	1	2	2	3	4	4	5	5
.42	2630	2636	2642	2649	2655	2661	2667	2673	2679	2685	1	1	2	2	3	4	4	5	6
.43	2692	2698	2704	2710	2716	2723	2729	2735	2742	2748	1	1	2	3	3	4	4	5	6
.44	2754	2761	2767	2773	2780	2786	2793	2799	2805	2812	1	1	2	3	3	4	4	5	6
.45	2818	2825	2831	2838	2844	2851	2858	2864	2871	2877	1	1	2	3	3	4	5	5	6
.46	2884	2891	2897	2904	2911	2917	2924	2931	2938	2944	1	1	2	3	3	4	5	5	6
.47	2951	2958	2965	2972	2979	2985	2992	2999	3006	3013	1	1	2	3	3	4	5	5	6
.48	3020	3027	3034	3041	3048	3055	3062	3069	3076	3083	1	1	2	3	4	4	5	6	6
.49	3090	3097	3105	3112	3119	3126	3133	3141	3148	3155	1	1	2	3	4	4	5	6	6

	0	1	2	3	4	5	6	7	8	9	平均差								
											1	2	3	4	5	6	7	8	9
·50	3162	3170	3177	3184	3192	3199	3206	3214	3221	3228	1	1	2	3	4	4	5	6	7
·51	3236	3243	3251	3258	3266	3273	3281	3289	3296	3304	1	2	2	3	4	5	5	6	7
·52	3311	3319	3327	3334	3342	3350	3357	3365	3373	3381	1	2	2	3	4	5	5	6	7
·53	3388	3396	3404	3412	3420	3428	3436	3443	3451	3459	1	2	2	3	4	5	6	6	7
·54	3467	3475	3483	3491	3499	3508	3516	3524	3532	3540	1	2	2	3	4	5	6	6	7
·55	3548	3556	3565	3573	3581	3589	3597	3606	3614	3622	1	2	2	3	4	5	6	7	7
·56	3631	3639	3648	3656	3664	3673	3681	3690	3698	3707	1	2	3	3	4	5	6	7	8
·57	3715	3724	3733	3741	3750	3758	3767	3776	3784	3793	1	2	3	3	4	5	6	7	8
·58	3802	3811	3819	3828	3837	3846	3855	3864	3873	3882	1	2	3	4	4	5	6	7	8
·59	3890	3899	3908	3917	3926	3936	3945	3954	3963	3972	1	2	3	4	5	5	6	7	8
·60	3981	3990	3999	4009	4018	4027	4036	4046	4055	4064	1	2	3	4	5	6	6	7	8
·61	4074	4083	4093	4102	4111	4121	4130	4140	4150	4159	1	2	3	4	5	6	7	8	9
·62	4169	4178	4188	4198	4207	4217	4227	4236	4246	4256	1	2	3	4	5	6	7	8	9
·63	4266	4276	4285	4295	4305	4315	4325	4335	4345	4355	1	2	3	4	5	6	7	8	9
·64	4365	4375	4385	4395	4406	4416	4426	4436	4446	4457	1	2	3	4	5	6	7	8	9
·65	4467	4477	4487	4498	4508	4519	4529	4539	4550	4560	1	2	3	4	5	6	7	8	9
·66	4571	4581	4592	4603	4613	4624	4634	4645	4656	4667	1	2	3	4	5	6	7	9	10
·67	4677	4688	4699	4710	4721	4732	4742	4753	4764	4775	1	2	3	4	5	7	8	9	10
·68	4786	4797	4808	4819	4831	4842	4853	4864	4875	4887	1	2	3	4	6	7	8	9	10
·69	4898	4909	4920	4932	4943	4955	4966	4977	4989	5000	1	2	3	5	6	7	8	9	10
·70	5012	5023	5035	5047	5058	5070	5082	5093	5105	5117	1	2	4	5	6	7	8	9	11
·71	5129	5140	5152	5164	5176	5188	5200	5212	5224	5236	1	2	4	5	6	7	8	10	11
·72	5248	5260	5272	5284	5297	5309	5321	5333	5346	5358	1	2	4	5	6	7	9	10	11
·73	5370	5383	5395	5408	5420	5433	5445	5458	5470	5483	1	3	4	5	6	8	9	10	11
·74	5495	5508	5521	5534	5546	5559	5572	5585	5598	5610	1	3	4	5	6	8	9	10	12
·75	5623	5636	5649	5662	5675	5689	5702	5715	5728	5741	1	3	4	5	7	8	9	10	12
·76	5754	5768	5781	5794	5808	5821	5834	5848	5861	5875	1	3	4	5	7	8	9	11	12
·77	5888	5902	5916	5929	5943	5957	5970	5984	5998	6012	1	3	4	5	7	8	10	11	12
·78	6026	6039	6053	6067	6081	6095	6109	6124	6138	6152	1	3	4	6	7	8	10	11	13
·79	6166	6180	6194	6209	6223	6237	6252	6266	6281	6295	1	3	4	6	7	9	10	11	13
·80	6310	6324	6339	6353	6368	6383	6397	6412	6427	6442	1	3	4	6	7	9	10	12	13
·81	6457	6471	6486	6501	6516	6531	6546	6561	6577	6592	2	3	5	6	8	9	11	12	14
·82	6607	6622	6637	6653	6668	6683	6699	6714	6730	6745	2	3	5	6	8	9	11	12	14
·83	6761	6776	6792	6808	6823	6839	6855	6871	6887	6902	2	3	5	6	8	9	11	13	14
·84	6918	6934	6950	6966	6982	6998	7015	7031	7047	7063	2	3	5	6	8	10	11	13	15
·85	7079	7096	7112	7129	7145	7161	7178	7194	7211	7228	2	3	5	7	8	10	12	13	15
·86	7244	7261	7278	7295	7311	7328	7345	7362	7379	7396	2	3	5	7	8	10	12	13	15
·87	7413	7430	7447	7464	7482	7499	7516	7534	7551	7568	2	3	5	7	9	10	12	14	16
·88	7586	7603	7621	7638	7656	7674	7691	7709	7727	7745	2	4	5	7	9	11	12	14	16
·89	7762	7780	7798	7816	7834	7852	7870	7889	7907	7925	2	4	5	7	9	11	13	14	16
·90	7943	7962	7980	7998	8017	8035	8054	8072	8091	8110	2	4	6	7	9	11	13	15	17
·91	8128	8147	8166	8185	8204	8222	8241	8260	8279	8299	2	4	6	8	9	11	13	15	17
·92	8318	8337	8356	8375	8395	8414	8433	8453	8472	8492	2	4	6	8	10	12	14	15	17
·93	8511	8531	8551	8570	8590	8610	8630	8650	8670	8690	2	4	6	8	10	12	14	16	18
·94	8710	8730	8750	8770	8790	8810	8831	8851	8872	8892	2	4	6	8	10	12	14	16	18
·95	8913	8933	8954	8974	8995	9016	9036	9057	9078	9099	2	4	6	8	10	12	15	17	19
·96	9120	9141	9162	9183	9204	9226	9247	9268	9290	9311	2	4	6	8	11	13	15	17	19
·97	9333	9354	9376	9397	9419	9441	9462	9484	9506	9528	2	4	7	9	11	13	15	17	20
·98	9550	9572	9594	9616	9638	9661	9683	9705	9727	9750	2	4	7	9	11	13	16	18	20
·99	9772	9795	9817	9840	9863	9886	9908	9931	9954	9977	2	5	7	9	11	14	16	18	20